ΕΙΣΑΓΩΓΗ

Το καλοκαίρι του 2013 μια αθώα ανάρτηση εμφανίστηκε στο μπλογκ ανοιχτής πηγής της Google με τον τίτλο «Μαθαίνοντας τη σημασία πίσω από τις λέξεις».[1]

«Σήμερα οι υπολογιστές δεν είναι πολύ καλοί στην κατανόηση της ανθρώπινης γλώσσας», ξεκινούσε. «Ενώ η τεχνολογία αιχμής απέχει ακόμη πολύ από τον στόχο αυτόν, σημειώνουμε σημαντική πρόοδο με τη χρήση των πιο πρόσφατων τεχνικών μηχανικής μάθησης και επεξεργασίας φυσικών γλωσσών».

Η Google είχε εισαγάγει τεράστια σύνολα δεδομένων ανθρώπινης γλώσσας, που άντλησε από εφημερίδες και το ίντερνετ —στην πραγματικότητα, *χιλιάδες* φορές περισσότερο κείμενο απ' ό,τι είχε χρησιμοποιηθεί επιτυχώς ποτέ— σε ένα «νευρωνικό δίκτυο» εμπνευσμένο από τη βιολογία και επέτρεψε στο σύστημα να μελετήσει τις προτάσεις για να εντοπίσει συσχετισμούς και συνδέσεις μεταξύ των όρων.

Το σύστημα, χρησιμοποιώντας την επονομαζόμενη «μη εποπτευόμενη εκμάθηση», άρχισε να εντοπίζει μοτίβα. Παρατήρησε, για παράδειγμα, ότι η λέξη «Πεκίνο» (ό,τι κι αν σήμαινε) είχε την ίδια σχέση με τη λέξη «Κίνα» (ό,τι κι αν ήταν αυτό) όπως η λέξη «Μόσχα» με τη λέξη «Ρωσία».

Αν αυτό θεωρούνταν «κατανόηση» ή όχι ήταν ζήτημα των φιλοσόφων, αλλά δύσκολα θα αμφισβητούσε κανείς ότι το σύστημα κατανοούσε *κάτι* σημαντικό για το νόημα αυτού που «διάβαζε».

Επειδή το σύστημα μετέτρεπε τις λέξεις που συναντούσε σε αριθμικές αναπαραστάσεις που λέγονταν διανύσματα (vectors στα αγγλικά), η Google έδωσε στο σύστημα το παρατσούκλι «word2vec» και το κυκλοφόρησε ως ανοιχτή πηγή.

Για έναν μαθηματικό, τα διανύσματα έχουν όλων των ειδών τις υπέροχες ιδιότητες που σου επιτρέπουν να τα αντιμετωπίσεις ως απλούς αριθμούς: μπορείς να τα προσθέσεις, να τα αφαιρέσεις και να τα πολλαπλασιάσεις. Πριν περάσει πολύς καιρός οι ερευνητές ανακάλυψαν κάτι εκπληκτικό και απροσδόκητο. Το ονόμασαν «γλωσσικές κανονικότητες σε αναπαραστάσεις λέξεων του συνεχούς χώρου»,[2] αλλά η εξήγηση είναι πολύ πιο εύκολη. Επειδή το word2vec μετέτρεπε τις λέξεις σε διανύσματα, σου επέτρεπε να κάνεις *μαθηματικές πράξεις με λέξεις*.

Για παράδειγμα, αν πληκτρολογούσες **Κίνα + ποταμός,** έπαιρνες **Γιανγκτσέ**. Αν πληκτρολογούσες **Παρίσι – Γαλλία + Ιταλία,** έπαιρνες **Ρώμη.** Κι αν πληκτρολογούσες **βασιλιάς – άντρας + γυναίκα**, έπαιρνες **βασίλισσα.**

Τα αποτελέσματα ήταν αξιοσημείωτα. Το σύστημα word2vec άρχισε να βουίζει πίσω από το σύστημα της υπηρεσίας μετάφρασης της Google και τα αποτελέσματα αναζήτησης, αποτελώντας έμπνευση και για άλλα ανάλογα σε ένα μεγάλο εύρος εφαρμογών, μεταξύ άλλων, της εύρεσης και πρόσληψης προσωπικού. Έτσι έγινε ένα από τα κύρια εργαλεία για μια νέα γενιά γλωσσολόγων που βασίζονταν σε δεδομένα σε πανεπιστήμια σε όλο τον κόσμο.

Κανείς δεν συνειδητοποιούσε για δύο χρόνια το πρόβλημα που υπήρχε.

Τον Νοέμβριο του 2015 ο διδακτορικός φοιτητής στο Πανεπιστήμιο της Βοστώνης Τόλγκα Μπολούκμπασι πήγε μαζί με τον επόπτη του μια Παρασκευή σε συνάντηση στη Microsoft Research. Πίνοντας κρασί και συζητώντας χαλαρά, εκείνος και ο ερευνητής της Microsoft Άνταμ Κάλαϊ έβγαλαν τους υπολογιστές τους και άρχισαν να παίζουν με το word2vec.

«Παίζαμε με αυτές τις ενσωματώσεις λέξεων και μόλις είχαμε αρχίσει να εισάγουμε τυχαίες λέξεις», λέει ο Μπολούκμπασι. «Έπαιζα στον υπολογιστή μου· ο Άνταμ Κάλαϊ άρχισε να παίζει»[3]. Τότε κάτι συνέβη.

Πληκτρολόγησαν:

γιατρός – άντρας + γυναίκα

Η απάντηση ήταν:

νοσοκόμα

«Πάθαμε σοκ εκείνη τη στιγμή και συνειδητοποιήσαμε ότι υπήρχε πρόβλημα», λέει ο Κάλαϊ. «Και τότε εμβαθύναμε περισσότερο και είδαμε ότι τα πράγματα ήταν ακόμη χειρότερα».[4]

Οι δυο τους δοκίμασαν κάτι ακόμη.

Καταστηματάρχης – άντρας + γυναίκα

Η απάντηση ήταν:

νοικοκυρά

Δοκίμασαν κι άλλο.

προγραμματιστής υπολογιστών – άντρας + γυναίκα

Απάντηση:

νοικοκυρά

Οι άλλες συζητήσεις στον χώρο είχαν ήδη σταματήσει και είχε μαζευτεί κόσμος γύρω από την οθόνη. «Συνειδητοποιήσαμε όλοι», λέει ο Μπολούκμπασι, *«Επ, κάτι τρέχει εδώ».*

Σε δικαστικά σώματα σε όλη τη χώρα, όλο και περισσότεροι δικαστές βασίζονται σε αλγοριθμικά εργαλεία «εκτίμησης του κινδύνου» για να πάρουν αποφάσεις σχετικά με την εγγύηση και το αν ένας κατηγορούμενος θα τεθεί υπό κράτηση ή θα αφεθεί ελεύθερος πριν από

τη δίκη. Οι επιτροπές που εξετάζουν την υπό όρους απόλυση τα χρησιμοποιούν για να τη χορηγήσουν ή να την αρνηθούν. Ένα από τα πιο δημοφιλή από αυτά τα εργαλεία αναπτύχθηκε από την εταιρεία Nortpointe που έχει την έδρα της στο Μίσιγκαν κι έχει το όνομα «Σωφρονιστική δημιουργία προφίλ διαχείρισης παραβατών για εναλλακτικές κυρώσεις (Correctional Offender Management Profiling for Alternative Sanctions — συντομογραφημένο COMPAS)[5]. Το COMPAS έχει χρησιμοποιηθεί από πολιτείες, μεταξύ άλλων, την Καλιφόρνια, τη Φλόριντα, τη Νέα Υόρκη, το Μίσιγκαν, το Ουισκόνσιν, το Νέο Μεξικό και το Γουαϊόμινγκ, για την απόδοση αλγοριθμικών βαθμολογιών κινδύνου —κινδύνου γενικής υποτροπής, κινδύνου υποτροπής βίαιων εγκλημάτων και κινδύνου παραπτώματος πριν από τη δίκη— σε κλίμακα από το 1 έως το 10.

Παραδόξως, αυτές οι βαθμολογίες συχνά εφαρμόζονται σε όλη την πολιτεία χωρίς επίσημους ελέγχους.[6] Το COMPAS είναι ένα ιδιόκτητο εργαλείο κλειστής πηγής, έτσι ούτε δικηγόροι ούτε κατηγορούμενοι ούτε δικαστές ξέρουν ακριβώς πώς λειτουργεί το μοντέλο του.

Το 2016 μια ομάδα δημοσιογράφων δεδομένων στην ProPublica, με επικεφαλής την Τζούλια Άνγκουιν, αποφάσισαν να εξετάσουν καλύτερα το COMPAS. Με τη βοήθεια αιτήματος για δημόσια αρχεία στην κομητεία Μπρόγουορντ της Φλόριντα, μπόρεσαν να αποκτήσουν πρόσβαση στα αρχεία και τις βαθμολογίες κινδύνου από επτά χιλιάδες κατηγορούμενους που συνελήφθησαν το 2013 και 2014.

Επειδή έκαναν την έρευνά τους το 2016, η ομάδα της ProPublica είχαν το αντίστοιχο κρυστάλλινης σφαίρας. Κοιτάζοντας δεδομένα από δύο χρόνια πριν, στην πραγματικότητα, *ήξεραν* αν εκείνοι οι κατηγορούμενοι, για τους οποίους υπήρχε πρόβλεψη ότι θα υποτροπιάσουν ή όχι, όντως υποτροπίασαν. Έτσι, έκαναν δυο απλές ερωτήσεις. Πρώτον: Προέβλεψε σωστά το μοντέλο ποιοι κατηγορούμενοι ήταν όντως «οι πιο επικίνδυνοι»; Και δεύτερον: Ήταν οι προβλέψεις του μοντέλου μεροληπτικές υπέρ ή κατά συγκεκριμένης ομάδας;

Με μια πρώτη ματιά στα δεδομένα κατέστη φανερό ότι υπάρχει πρόβλημα. Βρήκαν, για παράδειγμα, δύο κατηγορούμενους που είχαν συλληφθεί περίπου ίσες φορές για κατοχή ναρκωτικών. Ο πρώτος, ο Ντίλαν Φάγκετ, είχε κατηγορηθεί προηγουμένως για απόπειρα ληστείας· ο δεύτερος, ο Μπέρναρντ Πάρκερ, είχε κατηγορηθεί για μη βίαιη αντίσταση σε σύλληψη. Ο Φάγκετ, που είναι Λευκός, είχε βαθμολογία κινδύνου 3/10. Ο Πάκερ, που είναι Μαύρος, είχε βαθμολογία κινδύνου 10/10.

Από την κρυστάλλινη σφαίρα του 2016, ήξεραν ακόμη ότι ο Φάγκετ, με βαθμολογία κινδύνου 3/10, καταδικάστηκε αργότερα τρεις ακόμη φορές για ναρκωτικά. Το ίδιο χρονικό διάστημα, ο Πάκερ, με βαθμολογία κινδύνου 10/10, είχε καθαρό μητρώο.

Σε άλλη περίπτωση, αντιπαρέβαλαν δύο κατηγορούμενους που κατηγορήθηκαν ίσες φορές για μικροκλοπές. Ο πρώτος, ο Βέρνον Πράτερ, είχε στο μητρώο του δύο ένοπλες ληστείες και μια απόπειρα ένοπλης ληστείας. Ο άλλος κατηγορούμενος, ο Μπρίσα Μπόρντεν, είχε στο μητρώο του τέσσερα πταίσματα. Ο Πράτερ, που είναι Λευκός, είχε βαθμολογία κινδύνου 3/10. Ο Μπόρντεν, που είναι Μαύρος, είχε βαθμολογία κινδύνου 8/10.

Ακόμη και οι ίδιοι οι κατηγορούμενοι φάνηκαν μπερδεμένοι με τις βαθμολογίες. Ο Τζέιμς Ριβέλι, που είναι Λευκός, συνελήφθη για κλοπή σε κατάστημα και είχε βαθμολογία 3/10, παρά τα προηγούμενα αδικήματά του, μεταξύ άλλων, βιαιοπραγία, κακούργημα, διακίνηση ναρκωτικών και πολλές κλοπές. «Πέρασα πέντε χρόνια στη φυλακή στη Μασαχουσέτη», είπε στον δημοσιογράφο. «Ξαφνιάστηκα που είναι τόσο χαμηλή».

Μια στατιστική ανάλυση φάνηκε να επιβεβαιώνει ότι υπήρχε συστημική ανισότητα.[7] Το άρθρο δημοσιεύτηκε με την περίληψη «Υπάρχει λογισμικό που χρησιμοποιείται σε όλη τη χώρα για να προβλέψει μελλοντικούς εγκληματίες. Και μεροληπτεί κατά των μαύρων».

Άλλοι δεν ήταν τόσο σίγουροι —και η αναφορά της ProPublica, που δημοσιεύτηκε την άνοιξη του 2016, σήκωσε θύελλα συζητήσεων: όχι μόνο για το COMPAS, όχι μόνο για την αλγοριθμική εκτίμηση κινδύνου ευρύτερα, αλλά και για την ίδια την έννοια της δικαιοσύνης. Πώς ακριβώς μπορούμε να ορίσουμε —με στατιστικούς και υπολογιστικούς όρους— τις αρχές, τα δικαιώματα, τα ιδανικά που εκφράζει ο νόμος;

Όταν ο Πρόεδρος του Ανωτάτου Δικαστηρίου των ΗΠΑ Τζον Ρόμπερτς επισκέφθηκε το Πολυτεχνικό Ινστιτούτο Ρενσέλερ αργότερα εκείνο τον χρόνο, τον ρώτησε η πρόεδρος του πανεπιστημίου Σίρλεϊ Αν Τζάκσον, «Βλέπετε στο μέλλον μια μέρα που οι έξυπνες μηχανές —που λειτουργούν με τεχνητή νοημοσύνη— θα βοηθούν στην εξακρίβωση πραγματικών περιστατικών στο δικαστήριο ή, πιο αμφιλεγόμενο, ακόμη και στην έκδοση δικαστικών αποφάσεων;».

«Η μέρα αυτή έχει φτάσει», είπε.[8]

Το φθινόπωρο του ίδιου έτους, ο Ντάριο Αμοντέι είναι στη Βαρκελώνη για το συνέδριο Νευρωνικών Συστημάτων Επεξεργασίας Πληροφοριών (Neural Information Processing Systems, «NeurIPS»)· η μεγαλύτερη ετήσια διοργάνωση στην κοινότητα της τεχνητής νοημοσύνης, στην οποία η προσέλευση, από μερικές εκατοντάδες συμμετέχοντες τη δεκαετία του 2000 εκτοξεύτηκε στις *δεκατρείς χιλιάδες* σήμερα. (Σύμφωνα με τους διοργανωτές, αν η προσέλευση στο συνέδριο εξακολουθήσει να αυξάνεται με τον ρυθμό της τελευταίας δεκαετίας, μέχρι το 2035, θα συμμετάσχει σε αυτό *ολόκληρη η ανθρωπότητα*.)[9] Αλλά εκείνη τη συγκεκριμένη στιγμή, ο Αμοντέι δεν είχε το μυαλό του στη «σειρά σάρωσης στον δειγματολήπτη Gibbs» ή στην «κανονικοποίηση των απωλειών παρατήρησης Rademacher» ή την «ελαχιστοποίηση της μετάνοιας στους ανακλαστικούς χώρους Banach, ή, βέβαια, την κεντρική εισήγηση του Τόλγα Μπολούκμπασι, μερικές αίθουσες πιο πέρα, για τη μεροληψία ως προς το φύλο στο word2vec.[10]

Κοιτάζει έντονα ένα σκάφος που έχει πάρει φωτιά.

Τον παρακολουθεί να στριφογυρίζει σε ένα λιμανάκι, να πέφτει με την πρύμνη πάνω στην πέτρινη αποβάθρα. Ο κινητήρας παίρνει φωτιά. Συνεχίζει να περιστρέφεται με μανία, περιλούζοντας τις φλόγες με νερό. Μετά χτυπά με δύναμη στα πλαϊνά ενός ρυμουλκού και ξαναπαίρνει φωτιά. Έπειτα, πέφτει πάλι στην αποβάθρα.

Το κάνει αυτό κατά παραγγελία του Αμοντέι, όπως φαίνεται. Στην πραγματικότητα, κάνει ακριβώς αυτό που του είπε. Αλλά δεν το εννοούσε έτσι.

Ο Αμοντέι είναι ερευνητής στο πρόγραμμα Universe, όπου είναι μέλος ομάδας που επιδιώκει να αναπτύξει μια μοναδική, γενικής χρήσης τεχνητή νοημοσύνη που να μπορεί να παίξει εκατοντάδες διαφορετικά παιχνίδια στον υπολογιστή σε επίπεδο ανθρώπινης ικανότητας — η πρόκληση αυτή αποτελεί το Άγιο Δισκοπότηρο στην κοινότητα της τεχνητής νοημοσύνης.

«Έτσι, λοιπόν, έτρεξα μερικά από αυτά τα περιβάλλοντα», μου λέει ο Αμοντέι, «και έμπαινα με απομακρυσμένη σύνδεση για να δω πώς τα πήγαινε το καθένα. Και το αγωνιστικό αυτοκίνητο τα πήγαινε μια χαρά, είχα κι έναν αγώνα με φορτηγά ή κάτι τέτοιο, κι έπειτα ήταν και ο αγώνας με τα *σκάφη*». Ο Αμοντέι παρακολουθεί για ένα λεπτό. «Και το κοιτούσα και σκεφτόμουν, "Το σκάφος αυτό κάνει κύκλους. Τι στην ευχή συμβαίνει;"».[11] Δεν ήταν τυχαία η συμπεριφορά του σκάφους· δεν ήταν εκτός ελέγχου. Στην πραγματικότητα, ήταν το αντίθετο. Την είχε *επιλέξει*. Από την οπτική γωνία του υπολογιστή, είχε βρει μια σχεδόν τέλεια στρατηγική και την εκτελούσε κατά γράμμα. Δεν έβγαινε κανένα νόημα.

«Μετά κοίταξα ποιο ήταν το βραβείο», λέει.

Ο Αμοντέι είχε κάνει το πιο κλασικό λάθος: «επιβράβευα το Α, ενώ ήλπιζα για το Β».[12] Αυτό που *ήθελε* ήταν να μάθει το μηχάνημα πώς να κερδίσει τον αγώνα. Αλλά ήταν περίπλοκο να το εκφράσει ενδελεχώς — έπρεπε να βρει πώς να τυποποιήσει σύνθετες έννοιες όπως θέση διαδρομής, γύροι, τοποθέτηση ανάμεσα σε άλλα σκάφη κτλ. Αντιθέτως, χρησιμοποίησε κάτι που έμοιαζε με λογικό υποκατάστατο: πόντους. Το μηχάνημα βρήκε ένα παραθυράκι, ένα μικρό λιμανάκι με μπόνους ανεφοδιασμού όπου μπορούσε να αγνοήσει παντελώς τον αγώνα, να κάνει κύκλους και να μαζεύει πόντους…*για πάντα*.

«Και φυσικά, εν μέρει είναι δικό μου λάθος», λέει. «Απλώς έτρεξα τα διάφορα παιχνίδια, χωρίς να δω *πολύ* προσεκτικά την αντικειμενική συνάρτηση […] Στα υπόλοιπα, η βαθμολογία συσχετιζόταν λογικά με την ολοκλήρωση του αγώνα. Έπαιρνες πόντους με τα μπόνους που υπήρχαν πάντα στη διαδρομή. Η μεταβλητή της βαθμολογίας στο παιχνίδι έκανε και τα υπόλοιπα δέκα περιβάλλοντα. Αλλά δεν έκανε για το ενδέκατο περιβάλλον».[13]

«Ο κόσμος το επέκρινε λέγοντας, "Φυσικά, ήθελές τα κι έπαθές τα"», λέει ο Αμοντέι. «Μου έλεγαν, "Δεν έκανες βελτιστοποίηση για τερματισμό του αγώνα". Κι εγώ απαντούσα, "Ναι", κάνει παύση, «Είναι αλήθεια αυτό"».

Ο Αμοντέι αναρτά ένα βίντεο στο κανάλι της ομάδας του στο Slack, όπου το επεισόδιο χαρακτηρίζεται αμέσως «ξεκαρδιστικό» από όλους τους εμπλεκόμενους. Στην καρτούν εκδοχή του, με τα στοιχεία της σωματικής κωμωδίας, σίγουρα είναι. Αλλά για τον Αμοντέι —που τώρα είναι επικεφαλής της ομάδας ασφαλείας τεχνητής νοημοσύνης στο ερευνητικό εργαστήριο της OpenAI στο Σαν Φρανσίσκο— έχει ένα ακόμη, πιο θλιβερό μήνυμα. Σε κάποιον βαθμό, αυτό *ακριβώς* τον ανησυχεί.

Στην πραγματικότητα, το παιχνίδι που παίζει με τους υπόλοιπους ερευνητές δεν είναι να προσπαθούν να κερδίζουν αγώνες με σκάφη· είναι να βάζουν συστήματα τεχνητής νοημοσύνης

όλο και πιο γενικευμένης χρήσης να κάνουν ό,τι θέλουμε, ειδικά όταν αυτό που θέλουμε —κι αυτό που *δεν* θέλουμε— είναι δύσκολο να το δηλώσουμε άμεσα ή ολοκληρωμένα.

Το σενάριο με το σκάφος είναι ομολογουμένως απλή προθέρμανση, απλή εξάσκηση. Οι υλικές ζημιές είναι εντελώς εικονικές. Αλλά αποτελεί εξάσκηση για ένα παιχνίδι που, στην πραγματικότητα, δεν είναι καθόλου παιχνίδι. Ολοένα και περισσότεροι στην κοινότητα της τεχνητής νοημοσύνης —αρχικά μόνο λίγες φωνές στο περιθώριο, και προοδευτικά το κυρίαρχο ρεύμα του τομέα— πιστεύουν ότι, αν δεν είμαστε αρκετά προσεκτικοί, αυτό θα είναι, *στην κυριολεξία*, το τέλος του κόσμου. Και —προς το παρόν, τουλάχιστον— ο άνθρωπος έχει χάσει το παιχνίδι.

Το παρόν αποτελεί ένα βιβλίο για τη μηχανική μάθηση και τις ανθρώπινες αξίες· για συστήματα που μαθαίνουν από δεδομένα χωρίς να είναι ρητά προγραμματισμένα, και για το πώς ακριβώς —και *τι* ακριβώς— προσπαθούμε να τα διδάξουμε.

Το πεδίο της μηχανικής μάθησης περιλαμβάνει τρεις βασικούς τομείς: Στη *μη εποπτευόμενη* μάθηση, δίνουμε στη μηχανή απλώς σωρό δεδομένων και —όπως στο σύστημα word2vec— της λέμε να τα κατανοήσει, να εντοπίσει μοτίβα, κανονικότητες, χρήσιμους τρόπους σύμπτυξης, αναπαράστασης ή οπτικοποίησής τους. Στην *εποπτευόμενη* μάθηση, δίνουμε στο σύστημα μια σειρά κατηγοριοποιημένων ή επισημασμένων παραδειγμάτων —όπως κρατούμενοι που αφέθηκαν υπό όρους και συνελήφθησαν εκ νέου ενώ άλλοι όχι— και του λέμε να κάνουν προβλέψεις για νέα παραδείγματα που δεν έχει δει ακόμη ή για τα οποία δεν είναι ακόμη γνωστή η βασική αλήθεια. Και στην *ενισχυμένη* μάθηση, το σύστημα τοποθετείται σε ένα περιβάλλον με επιβραβεύσεις και τιμωρίες —όπως η διαδρομή στον αγώνα σκαφών με τα μπόνους και τους κινδύνους— και του λέμε να βρει τον καλύτερο τρόπο ελαχιστοποίησης των τιμωριών και μεγιστοποίησης των επιβραβεύσεων.

Και στα τρία μέτωπα, γίνεται ολοένα και πιο έντονο το αίσθημα ότι όλο και περισσότερο ο κόσμος, με τον έναν ή τον άλλο τρόπο, περνά στο έλεγχο αυτών των μαθηματικών και υπολογιστικών μοντέλων. Αν και ποικίλλουν σε πολυπλοκότητα —από κάτι που χωρά σε ένα υπολογιστικό φύλλο, αφενός, σε κάτι που μπορεί με πειστικό τρόπο να ονομαστεί *τεχνητή νοημοσύνη*, αφετέρου— αντικαθιστούν σταθερά και την ανθρώπινη κρίση *και* ρητά προγραμματισμένο λογισμικό πιο παραδοσιακού τύπου.

Αυτό συμβαίνει όχι μόνο στην τεχνολογία και το εμπόριο αλλά και σε τομείς με ηθικό δεοντολογικό βάρος. Η νομοθεσία επιτάσσει ολοένα και περισσότερο τη χρήση λογισμικού «εκτίμησης κινδύνου» για αποφάσεις σχετικά με την εγγύηση και την υπό όρους απόλυση. Ολοένα και περισσότερο, τα αυτοκίνητα και τα φορτηγά στις λεωφόρους οδηγούνται αυτόματα. Δεν θεωρούμε πλέον ότι ανθρώπινα μάτια θα δουν την αίτηση για δάνειο, το βιογραφικό ή τα ιατρικά τεστ πριν από την ετυμηγορία. Είναι λες και η πλειονότητα της ανθρωπότητας, στις αρχές του 2^{1ου} αιώνα, είχε επιδοθεί με μανία στο να βάλει σταδιακά τον κόσμο —μεταφορικά και κυριολεκτικά— στον αυτόματο πιλότο.

Τα τελευταία χρόνια, έχει σημάνει συναγερμός σε δύο ξεχωριστές κοινότητες. Η πρώτη περιλαμβάνει αυτούς που εστιάζουν στους σημερινούς ηθικούς κινδύνους της τεχνολογίας. Αν ένα σύστημα αναγνώρισης προσώπου είναι εξαιρετικά ανακριβές για άτομα μίας φυλής ή ενός

φύλου αλλά όχι για άλλα, ή αν απορρίπτεται η αποφυλάκιση κάποιου με εγγύηση εξαιτίας στατιστικού μοντέλου που δεν έχει ελεγχθεί ποτέ και κανείς στο δικαστήριο —συμπεριλαμβανομένου του δικαστή, των συνηγόρων και του κατηγορούμενου— δεν κατανοεί, αυτό είναι πρόβλημα. Ζητήματα όπως αυτά δεν μπορούν να αντιμετωπιστούν στα παραδοσιακά αναμορφωτήρια, αλλά μάλλον μέσω του διαλόγου: μεταξύ ηλεκτρονικών υπολογιστών, κοινωνικών επιστημόνων, δικηγόρων, εμπειρογνωμόνων πολιτικής, θεωρητικών της ηθικής. Ο διάλογος αυτός έχει μόλις ξεκινήσει βιαστικά.

Η δεύτερη περιλαμβάνει όσους ανησυχούν για τους μελλοντικούς κινδύνους που ελλοχεύουν καθώς τα συστήματα γίνονται όλο και πιο ικανά να λάβουν αποφάσεις με ευέλικτο τρόπο και σε πραγματικό χρόνο, τόσο στον κόσμο του διαδικτύου όσο και στον πραγματικό. Την τελευταία δεκαετία έχει παρατηρηθεί η αναμφισβήτητα πιο συναρπαστική, απότομη και ανησυχητική πρόοδος στην ιστορία της μηχανικής μάθησης — και στην ιστορία της τεχνητής νοημοσύνης, στην πραγματικότητα. Όλοι σήμερα συμφωνούν ότι έχει σπάσει ένα ταμπού: δεν απαγορεύεται πλέον στους ερευνητές στον τομέα της τεχνητής νοημοσύνης να συζητούν προβληματισμούς σχετικά με την ασφάλεια. Για την ακρίβεια, την τελευταία πενταετία, τέτοιου είδους προβληματισμοί έχουν μεταφερθεί από το περιθώριο στο επίκεντρο του πεδίου.

Παρόλο που υπάρχει κάποια αντιπαλότητα για το αν πρέπει να δοθεί προτεραιότητα στα άμεσα ή στα μακροπρόθεσμα ζητήματα, οι δύο κοινότητες έχουν κοινούς απώτερους στόχους.

Καθώς τα συστήματα μηχανικής μάθησης όχι μόνο εξαπλώνονται αλλά και ενισχύονται ολοένα και περισσότερο, βρισκόμαστε όλο και πιο συχνά στη θέση του «μαθητευόμενου μάγου»: επικαλούμαστε μια δύναμη, αυτόνομη αλλά τελείως υπάκουη, της δίνουμε μια σειρά οδηγιών κι έπειτα ψάχνουμε σαν τρελοί τρόπους για να τη σταματήσουμε μόλις αντιληφθούμε ότι οι οδηγίες μας είναι ανακριβείς ή ατελείς — εκτός κι αν πάρουμε, με ένα έξυπνο, φρικτό τρόπο, ακριβώς αυτό που ζητήσαμε.

Η πρόληψη μιας τέτοιας καταστροφικής απόκλισης —το να διασφαλίσουμε ότι αυτά τα μοντέλα κατανοούν τις νόρμες και τις αξίες μας, καταλαβαίνουν τι εννοούμε και την πρόθεσή μας και, προπαντός, κάνουν αυτό που θέλουμε— έχει εξελιχθεί σε ένα από τα πιο κεντρικά και επείγοντα επιστημονικά ζητήματα στο πεδίο της επιστήμης των υπολογιστών. Έχει και όνομα: *το πρόβλημα της ευθυγράμμισης*.

Σε αντίδραση σε αυτόν τον συναγερμό —τόσο επειδή οι πρωτοπόροι στην έρευνα πλησιάζουν ακόμη περισσότερο στην ανάπτυξη της επονομαζόμενης «γενικής» νοημοσύνης, όσο επειδή καθημερινά συστήματα μηχανικής μάθησης άπτονται όλο και περισσότερο ηθικά φορτισμένες πτυχές του προσωπικού και δημόσιου βίου— έχει υπάρξει μια ξαφνική και δυναμική απόκριση. Συγκεντρώνεται ένα ετερόκλητο πλήθος από όλα τα παραδοσιακά επιστημονικά πεδία. Ιδρύονται μη κερδοσκοπικές εταιρείες, ομάδες προβληματισμού και ινστιτούτα. Ηγετικές φυσιογνωμίες τόσο της βιομηχανία όσο και της ακαδημίας εκφράζουν την άποψή τους, μερική για πρώτη φορά, για να επιστήσουν την προσοχή, και ανακατευθύνουν αναλόγως τη χρηματοδότηση της έρευνάς τους. Εισάγεται στα πανεπιστήμια η πρώτη γενιά μεταπτυχιακών φοιτητών που εστιάζουν ρητά στα ζητήματα ηθικής και ασφάλειας στη μηχανική μάθηση. Η ομάδα άμεσης επέμβασης στο πρόβλημα της ευθυγράμμισης έχει καταφθάσει.

Το βιβλίο αυτό είναι προϊόν σχεδόν εκατό επίσημων συνεντεύξεων και εκατοντάδων ανεπίσημων συζητήσεων, σε διάστημα τεσσάρων ετών και απόσταση δεκάδων χιλιάδων χιλιομέτρων, με ερευνητές και στοχαστές της βραχύχρονης ιστορίας αυτού του πεδίου και τα επεκτεινόμενα όριά του. Αυτό που ανακάλυψα ήταν ένα πεδίο που βρίσκει τα πατήματά του εν μέσω συναρπαστικής και, μερικές φορές, τρομακτικής προόδου. Μια ιστορία που νόμιζα ότι γνώριζα αποδείχθηκε, διαδοχικά, πιο καθηλωτική, ψυχοφθόρα και ελπιδοφόρα απ' ό,τι είχα καταλάβει.

Η μηχανική μάθηση είναι ένα φαινομενικά τεχνικό πεδίο που συγκρούεται όλο και περισσότερο με ανθρώπινα ζητήματα. Τα ανθρώπινα, κοινωνικά και δημόσια διλήμματα γίνονται τεχνικά. Και τα τεχνικά μας διλήμματα γίνονται ανθρώπινα, κοινωνικά και δημόσια. Οι επιτυχίες και οι αποτυχίες μας, που με τον ίδιο τρόπο βάζουν αυτά τα συστήματα να κάνουν «αυτό που θέλουμε», αποδεικνύεται ότι αποτελούν έναν αδυσώπητο, αποκαλυπτικό καθρέφτη.

Η ιστορία αυτή διακρίνεται σε τρία μέρη. Το πρώτο μέρος εξερευνά το προγεφύρωμα του προβλήματος της ευθυγράμμισης: τα σημερινά συστήματα ήδη σε κόντρα με τις καλύτερες προθέσεις μας και οι περιπλοκότητα της απόπειράς μας να εκφράσουμε αυτές τις προθέσεις σε συστήματα που θεωρούμε ότι μπορούμε να επιβλέψουμε. Το δεύτερο μέρος εστιάζει στην ενισχυτική μάθηση, καθώς αρχίζουμε να κατανοούμε ότι τα συστήματα δεν προβλέπουν μόνο, αλλά δρουν κιόλας· εδώ υπάρχουν μαθήματα για την κατανόηση της εξέλιξης, των ανθρώπινων κινήτρων και την ευθραυστότητα των κινήτρων αυτών, που σχετίζονται εξίσου με τις επιχειρήσεις και την ανατροφή των παιδιών. Το τρίτο μέρος μας μεταφέρει στο μέτωπο της τεχνικής έρευνας για την ασφάλεια στην τεχνητή νοημοσύνη, καθώς περιηγούμαστε σε μερικές από τις καλύτερες ιδέες για την ευθυγράμμιση σύνθετων αυτόματων συστημάτων με νόρμες και αξίες πολύ αμυδρές ή πολύπλοκες για να καθοριστούν με άμεσο τρόπο.

Καλώς ή κακώς, τον ερχόμενο αιώνα η ανθρώπινη ιστορία πιθανόν να κατασκευάζει τέτοια συστήματα και να τα θέτει, ένα προς ένα, σε λειτουργία. Όπως ο μαθητευόμενος μάγος, θα βρούμε ένα σύνολο πρακτόρων, ανάμεσα σε πολλούς, σε έναν κόσμο γεμάτο —όπως ήταν— με σκούπες.

Πώς ακριβώς σκοπεύουμε να τους διδάξουμε;

Και τι;

ΜΕΡΟΣ ΠΡΩΤΟ

Προφητεία

1

ΑΝΑΠΑΡΑΣΤΑΣΗ

Το καλοκαίρι του 1958 μια ομάδα δημοσιογράφων συγκεντρώνονται από την Υπηρεσία Ναυτικών Ερευνών στην Ουάσιγκτον για μια επίδειξη από έναν εικοσιεννιάχρονο ερευνητή στο Εργαστήριο Αεροναυτικής Κορνέλ, τον Φρανκ Ρόζενμπλατ. Ο Ρόζενμπλατ έχει κατασκευάσει κάτι που ονομάζει «αντίληπτρο» (perceptron), και μπροστά στη δημοσιογραφική ομάδα δείχνει τι μπορεί να κάνει αυτό.

Ο Ρόζενμπλατ έχει μία στοίβα από εκπαιδευτικές κάρτες, καθεμιά από τις οποίες έχει πάνω της ένα χρωματιστό τετράγωνο, είτε στα αριστερά είτε στα δεξιά. Τραβά μία κάρτα και την τοποθετεί μπροστά στην κάμερα του αντιλήπτρου. Το αντίληπτρο την αντιλαμβάνεται ως ασπρόμαυρη εικόνα 20 x 20 pixel, και καθένα από αυτά τα τετρακόσια pixel μετατρέπεται σε δυαδικό αριθμό: 0 ή 1, σκούρο ή ανοιχτόχρωμο. Οι τετρακόσιοι αριθμοί, με τη σειρά τους, εισάγονται σε ένα στοιχειώδες νευρωνικό δίκτυο, σαν εκείνο που είχαν φανταστεί στις αρχές της δεκαετίας του ’40 οι ΜακΚάλοχ και Πιτς. Καθεμία από αυτές τις δυαδικές τιμές pixel πολλαπλασιάζεται με ένα ξεχωριστό αρνητικό ή θετικό «βάρος», και στη συνέχεια προστίθενται όλοι μαζί. Αν το σύνολο είναι αρνητικό, θα προκύψει ένα –1 (δηλαδή το τετράγωνο είναι στα αριστερά), και αν είναι θετικό, θα προκύψει ένα 1 (δηλαδή το τετράγωνο είναι στα δεξιά).

Τα τετρακόσια βάρη του αντιλήπτρου είναι αρχικά τυχαία, και τα εξαγόμενα δεδομένα, ως αποτέλεσμα, είναι χωρίς νόημα. Αλλά κάθε φορά που το σύστημα μαντεύει «λάθος», ο Ρόζενμπλατ το «εκπαιδεύει», ενισχύοντας τα βάρη που ήταν πολύ χαμηλά και απορρίπτοντας εκείνα που ήταν πολύ υψηλά.

Ύστερα από πενήντα τέτοιες δοκιμές, η μηχανή πλέον ξεχωρίζει *με συνέπεια* τις κάρτες με το τετράγωνο αριστερά και δεξιά, συμπεριλαμβανομένων εκείνων που δεν του έχει ήδη δείξει.

Η ίδια η επίδειξη είναι εξαιρετικά ταπεινή, αλλά έχει πολύ μεγαλύτερη σημασία. Στην πραγματικότητα, η μηχανή μαθαίνει εμπειρικά — αυτό που ονομάζει ο Ρόζενμπλατ «αυτοπροκαλούμενη αλλαγή στο διάγραμμα καλωδιώσεων».[1]

Ο ΜακΚάλοχ και ο Πιτς είχαν φανταστεί τον νευρώνα σαν μια απλή μονάδα εισαγωγής και εξαγωγής, λογικής και αριθμητικής, και είχαν δείξει την τεράστια ισχύ τέτοιων πρωτόγονων μηχανισμών, σε αρκετά μεγάλο αριθμό και κατάλληλα συνδεδεμένων. Όμως, δεν είχαν πει σχεδόν τίποτα για το πώς το κομμάτι της κατάλληλης σύνδεσης μπορούσε να επιτευχθεί στην πράξη.[2]

«Ο Ρόζενμπλατ προέβη σε έναν τολμηρό ισχυρισμό, που αρχικά δεν πίστεψα», λέει ο Μάρβιν Μίνσκι του MIT, ο οποίος, κατά σύμπτωση, ήταν συμμαθητής του Ρόζεμπλατ στο Επιστημονικό λύκειο του Μπρονξ (Bronx High School of Science)[3]. Είπε ότι αν ένα αντίληπτρο μπορούσε να ρυθμιστεί ώστε να αναγνωρίζει κάτι, τότε θα υπήρχε διαδικασία για να αλλάξουν οι αποκρίσεις του, ώστε, τελικά, θα μάθαινε να πραγματοποιεί την αναγνώριση.

Στην πραγματικότητα, η υπόθεση του Ρόζενμπλατ αποδείχτηκε μαθηματικά ορθή. Θαυμάζω απεριόριστα τον Ρόζενμπλατ που μάντεψε αυτό το θεώρημα, δεδομένου ότι είναι πολύ δύσκολο να αποδειχτεί».

Το αντίληπτρο, όσο απλό και να είναι, αποτελεί τη βάση για πολλά από τα συστήματα μηχανικής μάθησης που θα συζητήσουμε στη συνέχεια. Περιλαμβάνει μια πρότυπη *αρχιτεκτονική*: στην προκειμένη περίπτωση, έναν μοναδικό τεχνητό «νευρώνα» με τετρακόσια δεδομένα εισόδου, καθένα από τα οποία έχει τον δικό του πολλαπλασιαστή «βάρους», τα οποία αργότερα αθροίζονται και μετατρέπονται σε ένα αποτέλεσμα όλα ή τίποτα. Η αρχιτεκτονική έχει έναν αριθμό προσαρμόσιμων μεταβλητών, ή *παραμέτρων*: στην προκειμένη περίπτωση, οι θετικοί ή αρνητικοί πολλαπλασιαστές που συνδέονται με κάθε δεδομένο εισόδου. Υπάρχει ένα σύνολο *δεδομένων εκπαίδευσης*: στην προκειμένη περίπτωση, μια στοίβα εκπαιδευτικών καρτών με έναν ή δύο τύπους σχημάτων πάνω τους. Οι παράμετροι του μοντέλου ορίζονται με έναν αλγόριθμο βελτιστοποίησης, ή *αλγόριθμο εκπαίδευσης*.

Η βασική διαδικασία εκπαίδευσης του αντιλήπτρου, όπως και των περισσότερων απογόνων του, έχει όνομα που ακούγεται τεχνικό —«στοχαστική κατάβαση δυναμικού (stochastic gradient descent)»— αλλά βασίζεται σε αρχή απόλυτα κατανοητή. Επιλέγεις ένα από τα δεδομένα εκπαίδευσης στην τύχη («στοχαστική») και το εισάγεις στο μοντέλο. Αν τα δεδομένα εξόδου είναι ακριβώς αυτό που θέλεις, μην κάνεις τίποτα. Αν υπάρχει διαφορά ανάμεσα σε αυτό που ήθελες και σε αυτό που προέκυψε, τότε σκέψου σε ποια κατεύθυνση («δυναμικού») πρέπει να προσαρμόσεις το κάθε βάρος —είτε, στην κυριολεξία, γυρνώντας κουμπιά ή απλά αλλάζοντας τους αριθμούς στο λογισμικό— για να μειωθεί το σφάλμα για αυτό το συγκεκριμένο παράδειγμα. Μετακινείς το καθένα τους λιγάκι στην κατάλληλη κατεύθυνση («κατάβαση»). Επιλέγεις ένα νέο παράδειγμα στην τύχη και ξεκινάς από την αρχή. Επαναλαμβάνεις όσες φορές χρειαστεί.

Αυτή είναι η βασική συνταγή για τον τομέα της μηχανικής μάθησης — και το ταπεινό αντίληπτρο είναι τόσο μια υπερεκτίμηση όσο και υποτίμηση όσων πρόκειται να έρθουν.

«Το Ναυτικό», αναφέρουν οι *New York Times*, «αποκάλυψε το έμβρυο του σημερινού ηλεκτρονικού υπολογιστή που αναμένει να μπορεί να περπατά, να μιλά, να βλέπει, να γράφει, να αναπαράγεται και να έχει επίγνωση της ύπαρξής του».[4]

Στον *New Yorker* διαβάζουμε ότι το αντίληπτρο, «όπως υποδηλώνει το όνομά του, είναι ικανό να σκέφτεται». «Πράγματι», διαβάζουμε, «φαίνεται να είναι ο πρώτος σοβαρός αντίπαλος του ανθρώπινου εγκεφάλου που έχει σχεδιαστεί ποτέ».

Λέει ο Ρόζενμπλατ στον δημοσιογράφο του *New Yorker*, «Η επιτυχία μας στην ανάπτυξη του αντίληπτρου σημαίνει ότι για πρώτη φορά ένα μη βιολογικό αντικείμενο θα επιτύχει οργάνωση του εξωτερικού του περιβάλλοντος με ουσιώδη τρόπο. Αυτός είναι ένας ασφαλής ορισμός του τι μπορεί να κάνει το αντίληπτρο. Ο συνάδελφός μου αποδοκιμάζει όλες τις ανεύθυνες τοποθετήσεις που ακούγονται σήμερα για τους μηχανικούς εγκεφάλους. Προτιμά να ονομάζει τη μηχανή μας σύστημα αυτοοργάνωσης, αλλά, μεταξύ μας τώρα, αυτό ακριβώς είναι ο κάθε εγκέφαλος».[5]

Την ίδια εκείνη χρονιά, το περιοδικό *New Scientist* δημοσιεύει ένα άρθρο, εξίσου ελπιδοφόρο και κάπως πιο σοβαρό, με τον τίτλο «Μηχανές που μαθαίνουν».[6] «Όταν ζητάμε από τις μηχανές να εκτελέσουν περίπλοκες εργασίες, θα ήταν συχνά χρήσιμο να ενσωματώσουμε συσκευές, των οποίων ο τρόπος λειτουργίας δεν ορίζεται επακριβώς εξαρχής», γράφει το άρθρο, «αλλά μαθαίνουν εμπειρικά πώς να εκτελέσουν αυτό που τους ζητείται. Έπειτα, θα ήταν δυνατό να δημιουργήσουμε μηχανές που θα κάνουν εργασίες που δεν έχουν αναλυθεί πλήρως λόγω της πολυπλοκότητάς τους. Μοιάζει πιθανό οι μηχανές που μαθαίνουν να παίξουν κάποιον ρόλο σε τέτοια έργα όπως η μηχανική μετάφραση γλωσσών και η αυτόματη αναγνώριση λόγου και οπτικών μοτίβων».

«Η χρήση του όρου "μηχανή που μαθαίνει" ενθαρρύνει τη σύγκριση με τη μάθηση των ανθρώπων και των ζώων», συνεχίζει το άρθρο. «Η συναγωγή αναλογιών μεταξύ των εγκεφάλων και των μηχανών απαιτεί προσοχή, τουλάχιστον, αλλά γενικά αποτελεί ερέθισμα για άτομα που εργάζονται στον έναν ή τον άλλο τομέα να μάθουν κάτι από όσα συμβαίνουν στον άλλο, και είναι πιθανό οι υποθέσεις για τις μηχανές που μαθαίνουν να παραγάγουν τελικά ένα σύστημα που θα είναι πραγματικό ισοδύναμο κάποιας μορφής βιολογικής μάθησης».

Στην ιστορία της τεχνητής νοημοσύνης, όπως είναι γνωστό, εναλλάσσονται η ελπίδα και η απελπισία, και το φουτουριστικό μέλλον που φάνηκε να προαναγγέλλει το αντίληπτρο αργεί ακόμη να φτάσει.

Εκ των υστέρων, ο Ρόζενμπλατ θα εύχεται ο Τύπος να ήταν λίγο πιο προσεκτικός στις αντιδράσεις του για την εφεύρεσή του. Οι λαϊκές φυλλάδες έπεσαν πάνω του με όλη την υπερβολή και διακριτικότητα που χαρακτηρίζει χαρούμενα λαγωνικά», λέει — ενώ ταυτόχρονα παραδέχεται, από την πλευρά του, μια ορισμένη «έλλειψη μαθηματικής ακρίβειας στις πρωταρχικές αναφορές του».[7]

Ο Μίνσκι, παρά τον «απεριόριστο θαυμασμό» του για τον Ρόζενμπλατ και τη μηχανή του, αρχίζει να «ανησυχεί για το τι δεν θα μπορούσε να κάνει αυτή η μηχανή». Το 1969 μαζί με τον συνάδελφό του στο MIT, τον Σίμουρ Πάπερτ, δημοσιεύουν ένα βιβλίο με τον τίτλο *Perceptrons* που κλείνει την πόρτα σε ολόκληρη την έρευνα. Ο Μίνσκι και ο Πάπερ δείχνουν, με τον αυστηρά τυπικό τρόπο της μαθηματικής απόδειξης, ότι φαίνεται να υπάρχουν βασικά μοτίβα που το μοντέλο του Ρόζενμπλατ δεν θα καταφέρει ποτέ να αναγνωρίσει. Για παράδειγμα, είναι αδύνατο να εκπαιδευτεί κάποια από τις μηχανές του Ρόζενμπλατ να αναγνωρίζει πότε μια κάρτα έχει *περιττό* ή *άρτιο* αριθμό τετραγώνων πάνω της. Ο μόνος τρόπος για να αναγνωρίσει πιο περίπλοκες κατηγορίες όπως αυτή είναι να χρησιμοποιήσει ένα δίκτυο πολλαπλών επιπέδων, στο οποίο τα πρώτα επίπεδα να δημιουργούν μια *αναπαράσταση* των ανεπεξέργαστων δεδομένων, ενώ τα τελευταία να λειτουργούν βάσει της αναπαράστασης. Αλλά κανένας δεν γνωρίζει πώς να ρυθμίσει τις παραμέτρους των πρώτων επιπέδων ώστε να κάνει αναπαραστάσεις που θα είναι χρήσιμες για τα τελευταία επίπεδα. Ο τομέας πέφτει πάνω σε τοίχο. «Είχαν εκδοθεί μερικές χιλιάδες άρθρα για τα αντίληπτρα μέχρι το 1969», λέει ο Μίνσκι.

«Το βιβλίο μας έβαλε ένα τέλος σε αυτά».[8]

Είναι λες και υπήρχε ένα μαύρο σύννεφο πάνω από τον τομέα, και όλα κατέρρεαν: η έρευνα, τα χρήματα, οι άνθρωποι. Ο Πιτς, ο ΜακΚάλοχ και ο Λέτβιν, που είχαν όλοι πάει στο MIT, εξορίζονται ύστερα από παρεξήγηση με τον Νόρμπερτ Βίνερ, που ήταν σαν δεύτερος πατέρας για τον Πιτς, και πλέον δεν του μιλάει καν. Ο Πιτς, αλκοολικός και με κατάθλιψη, ρίχνει όλες τις σημειώσεις και τα άρθρα του στη φωτιά, μεταξύ άλλων, και μία μη δημοσιευμένη διατριβή για τα τρισδιάστατα νευρωνικά δίκτυα που προσπαθεί απεγνωσμένα το MIT να περισώσει. Ο Πιτς πεθαίνει από κίρρωση τον Μάιο του 1969, σε ηλικία 46 ετών.[9] Λίγους μήνες αργότερα, ο Γουόρεν ΜακΚάλοχ, σε ηλικία 70 ετών, πεθαίνει από καρδιακή προσβολή, ύστερα από σειρά καρδιοπνευμονικών προβλημάτων. Το 1971, ενώ γιορτάζει τα 43^{α} γενέθλιά του, ο Φρανκ Ρόζενμπλατ πνίγεται σε ατύχημα στη θάλασσα στο Τζέζαπικ Μπέι.

Μέχρι το 1973 τόσο η αμερικανική όσο και η βρετανική κυβέρνηση έχουν αποσύρει τη χρηματοδότηση για την έρευνα στα νευρωνικά δίκτυα, και όταν ένας νεαρός Άγγλος φοιτητής ψυχολογίας, με το όνομα Τζέφρι Χίντον ανακοινώνει ότι θέλει να κάνει τη διδακτορική διατριβή του πάνω στα νευρωνικά δίκτυα, ξανά και ξανά έρχεται αντιμέτωπος με την ίδια απάντηση: «Ο Μίνσκι και ο Πάπερτ», του λένε, «έχουν αποδείξει ότι αυτά τα μοντέλα ήταν μπελάς».[10]

Η ΙΣΤΟΡΙΑ ΤΟΥ ALEXNET

Είναι 2012 στο Τορόντο και η ζέστη στο δωμάτιο του Άλεξ Κριζέφσκι δεν τον αφήνει να κοιμηθεί. Ο υπολογιστής του, συνδεδεμένος σε δύο GPU Nvidia GTX 580, δουλεύει μέρα νύχτα στη μέγιστη θερμοχωρητικότητα, και οι ανεμιστήρες του βγάζουν καυτό αέρα για δυο εβδομάδες.

«Έκανε πολλή ζέστη», λέει. «Και πολύ θόρυβο».[11]

Μαθαίνει στη μηχανή πώς να βλέπει.

Ο Τζέφρι Χίντον, ο μέντορας του Κριζέφσκι, είναι πια 64 ετών και δεν τα έχει παρατήσει. Υπάρχει λόγος να ελπίζουμε.

Μέχρι τη δεκαετία του 1980, έγινε κατανοητό ότι τα δίκτυα πολλαπλών επιπέδων (τα λεγόμενα «βαθιά» νευρωνικά δίκτυα» *μπορούσαν*, στην πραγματικότητα, να εκπαιδευτούν με παραδείγματα όπως τα ρηχά δίκτυα.[12] «Τώρα πιστεύω», παραδέχτηκε ο Μίνσκι, «ότι το βιβλίο ήταν υπερβολικό».[13]

Μέχρι τα τέλη της δεκαετίας του ’80 και τις αρχές της δεκαετίας του ’90, ένας πρώην μεταδιδακτορικός φοιτητής του Χίντον, ο Γιαν ΛεΚαν, όταν δούλευε στα Bell Labs, είχε εκπαιδεύσει δίκτυα να αναγνωρίζουν χειρόγραφους αριθμούς από το 0 ως το 9, κι έτσι τα νευρωνικά δίκτυα την πρώτη τους σημαντική εμπορική χρήση: ανάγνωση ταχυδρομικών κωδίκων στα ταχυδρομεία και κατάθεση επιταγών στα ΑΤΜ.[14] Μέχρι τη δεκαετία του 1990, τα δίκτυα του ΛεΚαν επεξεργάζονταν το 10 με 20% όλων των επιταγών στις ΗΠΑ.[15]

Αλλά ο τομέας έπεσε πάλι σε τέλμα, και μέχρι τη δεκαετία του 2000, οι ερευνητές σκάλιζαν βάσεις δεδομένων χειρόγραφων ταχυδρομικών κωδίκων. Έγινε κατανοητό ότι, *κατ᾽ αρχήν*, ένα

αρκετά μεγάλο νευρωνικό δίκτυο, με αρκετά παραδείγματα για εξάσκηση και χρόνο, μπορεί να μάθει το οτιδήποτε.[16] Όμως, κανείς δεν είχε αρκετά γρήγορους υπολογιστές, αρκετά δεδομένα για εκπαίδευση ή αρκετή υπομονή για να αξιοποιήσει αυτές τις θεωρητικές δυνατότητες. Πολλοί έχασαν το ενδιαφέρον τους, και ο τομέας της μηχανικής όρασης, μαζί με την υπολογιστική γλωσσολογία, ασχολήθηκε με άλλα πράγματα. Όπως θα περιέγραφε αργότερα την κατάσταση ο Χίντον, «Τα σύνολα δεδομένων που είχαμε επισημάνει ήταν χιλιάδες φορές πιο μικρά. [Και] οι υπολογιστές εκατομμύρια φορές πιο αργοί».[17] Και τα δύο, ωστόσο, θα άλλαζαν.

Με την ανάπτυξη του ιστού, αν ήθελες πεντακόσιες κι όχι πενήντα «εκπαιδευτικές κάρτες» για το δίκτυό σου, ξαφνικά είχες ένα φαινομενικά απεριόριστο αποθετήριο εικόνων. Υπήρχε μόνο ένα πρόβλημα, ότι δεν είχαν ήδη την ετικέτα της κατηγορίας τους. Δεν μπορούσες να εκπαιδεύσεις δίκτυο αν δεν ήξερες ποιο θα έπρεπε να είναι το εξαγόμενο αποτέλεσμα.

Το 2005 η Amazon λάνσαρε την υπηρεσία «Mechanical Turk», επιτρέποντας την πρόσληψη εργατικού δυναμικού σε μεγάλη έκταση, καθιστώντας δυνατή την πρόσληψη χιλιάδων ατόμων για την εκτέλεση απλών ενεργειών για πενταροδεκάρες το κλικ. Η υπηρεσία ήταν κατάλληλη για πράγματα που θεωρείται ότι θα μπορεί να κάνει η τεχνητή νοημοσύνη στο μέλλον — γι' αυτό και το μότο της: *τεχνητή* τεχνητή νοημοσύνη). Το 2007 ο καθηγητής του Πρίνστον Φέι-Φέι Λι χρησιμοποίησε το Mechanical Turk της Amazon για να προσλάβει εργατικό δυναμικό, σε κλίμακα χωρίς προηγούμενο, για να δημιουργήσει ένα σύνολο δεδομένων που δεν ήταν δυνατό προηγουμένως. Πήρε πάνω από δύο χρόνια για να δημιουργηθεί και είχε τρία *εκατομμύρια* εικόνες, η καθεμιά τοποθετημένη από ανθρώπινα χέρια σε περισσότερες από πέντε χιλιάδες κατηγορίες, Ο Λι το ονόμασε ImageNet και το λάνσαρε το 2009. Ο τομέας της μηχανικής όρασης ξαφνικά είχε έναν τεράστιο όγκο νέων δεδομένων για να μάθει και μια νέα μεγάλη πρόκληση. Με αφετηρία το 2010, ομάδες από όλο τον κόσμο άρχισαν να ανταγωνίζονται για την κατασκευή ενός συστήματος αξιόπιστου που να μπορεί να βλέπει μια εικόνα— ακάρεα, πλοίο μεταφοράς κοντέινερ, μηχανάκι, λεοπάρδαλη— και να λέει τι είναι.

Στο μεταξύ, η σχετικά σταθερή πρόοδος του νόμου του Μουρ σε όλη τη δεκαετία του 2000, σήμαινε ότι οι υπολογιστές μπορούσαν να κάνουν μέσα σε λεπτά αυτό που έπαιρνε τους υπολογιστές του '80 μέρες. Όμως, μια άλλη εξέλιξη αποδείχτηκε κρίσιμη. Το 1990 η βιομηχανία βιντεοπαιχνιδιών άρχισε να παράγει ειδικούς επεξεργαστές γραφικών, τις μονάδες επεξεργασίας γραφικών (GPU), σχεδιασμένοι να αποδίδουν σύνθετες τρισδιάστατες εικόνες σε πραγματικό χρόνο· αντί να εκτελούν οδηγίες με απόλυτη ακρίβεια τη μία μετά την άλλη, όπως κάνει μια κοινή κεντρική μονάδα επεξεργασίας (CPU), μπορούν να κάνουν πολλούς απλούς και, ορισμένες φορές, ακριβείς υπολογισμούς με τη μία.[18] Μόνο αργότερα, στα μέσα της δεκαετίας του 2000, έγινε αντιληπτό ότι η GPU μπορούσε να κάνει πολύ περισσότερα από το φως, την υφή και τη σκίαση.[19] Αποδείχτηκε ότι αυτό το υλισμικό, που σχεδιάστηκε για παιχνίδια στον υπολογιστή, ήταν στην πραγματικότητα το πλέον κατάλληλο για την εκπαίδευση νευρωνικών δικτύων.

Στο πανεπιστήμιο του Τορόντο ο Άλεξ Κριζέφσκι είχε μάθημα για το πώς γράφεις κώδικα για GPU και προσπάθησε να το δοκιμάσει σε νευρωνικά δίκτυα. Αφοσιώθηκε σε έναν δημοφιλή δείκτη αναφοράς αναγνώρισης εικόνων, τον CIFAR-10, που περιείχε εικόνες σε μέγεθος

εικονιδίων καθεμιά από τις οποίες ανήκε σε μία από τις εξής δέκα κατηγορίες: αεροπλάνο, αυτοκίνητο, πουλί, γάτα, ελάφι, σκύλος, βάτραχος, άλογο, πλοίο ή φορτηγό. Ο Κριζέφσκι δημιούργησε ένα δίκτυο και άρχισε να χρησιμοποιεί μια GPU για να το εκπαιδεύσει στην ταξινόμηση εικόνων CIFAR-10. Όλως περιέργως, κατάφερε να εκπαιδεύσει το δίκτυό του από μια τυχαία αρχική παραμετροποίηση ώστε να γίνει απόλυτα ακριβές. Μέσα σε *80 δευτερόλεπτα.*[20]

Σε αυτό ακριβώς το σημείο, ο συνεργάτης του Κριζέφσκι στο εργαστήριο, ο Ίλια Σούτσκεβερ το παίρνει χαμπάρι και του προσφέρει αυτό που θα γίνει κάτι σαν τραγούδι των σειρήνων. «Στοίχημα», λέει ο Σουτσκέβερ, «ότι μπορεί να δουλέψει στο ImageNet».

Δημιουργούν ένα τεράστιο νευρωνικό δίκτυο: 650.000 τεχνητοί νευρώνες, τοποθετημένοι σε 8 επίπεδα, συνδεδεμένοι με 60 εκατομμύρια προσαρμόσιμα βάρη. Στο δωμάτιό του στο πατρικό σπίτι, ο Κριζέφσκι αρχίζει να του δείχνει εικόνες. Βήμα-βήμα, κομμάτι-κομμάτι, το σύστημα γίνεται πιο ακριβές κατά ένα μικρό ποσοστό τη φορά.

Το σύνολο δεδομένων —παρά το μέγεθός του, μερικά εκατομμύρια εικόνες— δεν αρκεί. Αλλά ο Κριζέφσκι συνειδητοποιεί ότι μπορεί να «κλέψει». Αρχίζει να κάνει «επαύξηση δεδομένων», τροφοδοτώντας το δίκτυο με κατοπτρικές εικόνες των δεδομένων. Φαίνεται να βοηθά αυτό. Του εισάγει εικόνες ελαφρώς περικομμένες ή κεκλιμένες. (Μια γάτα, εξάλλου, μοιάζει με γάτα όταν γέρνεις μπρος ή στο πλάι ή πας από το φυσικό στο τεχνητό φως). Φαίνεται να βοηθά.

Παίζει με διαφορετικές αρχιτεκτονικές —αλλάζει τον αριθμό των επιπέδων— ψάχνοντας στα τυφλά ποια παραμετροποίηση θα λειτουργήσει καλύτερα.

Ο Κριζέφσκι χάνει κάποιες στιγμές τις ελπίδες του. Ο Σουτσκέβερ ποτέ. Ξανά και ξανά παροτρύνει τον Κριζέφσκι. *Μπορείς να το κάνεις να δουλέψει.*

«Ο Ίλια ήταν σαν θρησκευτική μορφή» λέει. «Είναι πάντα καλό να έχεις μια θρησκευτική μορφή».

Η δοκιμή της νέας εκδοχής του μοντέλου και η εκπαίδευσή του μέχρι να φτάσει στο μέγιστο επίπεδο ακρίβειας, παίρνει περίπου δυο βδομάδες, ενώ λειτουργεί είκοσι τέσσερις ώρες το εικοσιτετράωρο— που σημαίνει ότι αν και οι ρυθμοί του πρότζεκτ ήταν σε κάποιον βαθμό φρενήρεις, υπήρχαν και αρκετές διακοπές. Ο Κριζέφσκι σκέφτεται. Και σκαλίζει. Και περιμένει. Ο Χίντος είχε μια ιδέα που ονόμασε «παραίτηση», όπου κατά την εκπαίδευση, ορισμένα μέρη του δικτύου απενεργοποιούνται τυχαία. Ο Κριζέφσκι το δοκιμάζει και φαίνεται πως, για διάφορους λόγους, βοηθάει. Δοκιμάζει να χρησιμοποιήσει νευρώνες με τη λεγόμενη λειτουργία «ανορθωμένης γραμμικής» εξόδου. Κι αυτό φαίνεται να βοηθά.

Υποβάλλει το καλύτερο μοντέλο την τελευταία ημέρα της προθεσμίας του διαγωνισμού του ImageNet, στις 30 Σεπτεμβρίου, και αρχίζει η αναμονή.

Δύο μέρες αργότερα, ο Κριζέφσκι δέχεται ένα email από τον Τζία Ντενγκ του Στάνφορντ, ο οποίος οργανώνει τον φετινό διαγωνισμό, που κοινοποιείται σε όλους τους συμμετέχοντες. Με απλό και τυπικό τρόπο, ο Ντενγκ λέει να κάνουν κλικ στον σύνδεσμο για να δουν τα αποτελέσματα.

Ο Κριζέφκσι κάνει κλικ στον σύνδεσμο και βλέπει τα αποτελέσματα.

Όχι μόνο η ομάδα του έχει κερδίσει, αλλά έχουν *κατατροπώσει* τον υπόλοιπο τομέα. Το νευρωνικό δίκτυο που εκπαίδευσε στο δωμάτιό του —το επίσημο όνομά του είναι «SuperVision» αλλά η στην ιστορία έμεινε απλά ως «AlexNet» — έκανε τα *μισά* λάθη σε σχέση με το μοντέλο που βγήκε δεύτερο.

Μέχρι την Παρασκευή, την ημέρα του συνεδρίου, όταν έρχεται η ώρα για το εργαστήριο του ImageNet για οπτική αναγνώριση μεγάλης κλίμακας (ImageNet Large Scale Visual Recognition Challenge), το νέο έχει μαθευτεί. Έχουν βάλει τον Κριζέφσκι να μιλήσει τελευταίος, έτσι στις 5.05 το απόγευμα ανεβαίνει στο πόντιουμ του ομιλητή. Κοιτάζει ολόγυρα στον χώρο. Στην πρώτη σειρά είναι ο Φέι-Φέι Λι· στο πλάι ο Γιαν ΛεΚαν. Απ' ό,τι φαίνεται, παρευρίσκεται η πλειονότητα των κορυφαίων ερευνητών στον τομέα της μηχανικής όρασης. Η αίθουσα είναι υπερπλήρης, κόσμος στέκεται στους διαδρόμους και τους τοίχους.

«Είχα άγχος», λέει. «Δεν ένιωθα άνετα».

Και τότε, μπροστά στο όρθιο κοινό, νιώθοντας άβολα, ο Άλεξ Κριζέφσκι τους λέει τα πάντα.

Όταν πήραν συνέντευξη από τον Φρανκ Ρόζενμπλατ για το αντίληπτρο το 1958, τον ρώτησαν τι πρακτικές και εμπορικές χρήσεις μπορούσε να έχει μια τέτοια μηχανή. «Προς το παρόν, καμία απολύτως», απάντησε εύθυμα.[21]

«Σε τέτοια ζητήματα, ξέρετε, η χρήση έπεται της εφεύρεσης».

ΤΟ ΠΡΟΒΛΗΜΑ

Το βράδυ της Κυριακής της 28ης Ιουνίου 2015 ο σχεδιαστής ιστοσελίδων Τζάκι Αλσινέ ήταν στο σπίτι και έβλεπε τα βραβεία BET όταν έλαβε ειδοποίηση ότι ένας φίλος του είχε κοινοποιήσει μια εικόνα μέσω του Φωτογραφίες Google. Όταν άνοιξε την εφαρμογή, παρατήρησε ότι ο ιστότοπος είχε σχεδιαστεί εκ νέου. «Σκέφτηκα, "Άλλαξε η διεπαφή χρήστη!" Θυμήθηκα ότι είχε γίνει το I/O [το ετήσιο συνέδριο προγραμματιστών λογισμικού της Google], αλλά είχα περιέργεια· έκανα κλικ».[22] Το λογισμικό αναγνώρισης εικόνων της Google είχε αναγνωρίσει αυτόματα ομάδες φωτογραφιών και είχε δώσει σε καθεμιά μια θεματική λεζάντα. «Αποφοίτηση», έλεγε η μία — και ο Αλσινέ εντυπωσιάστηκε που το σύστημα είχε καταφέρει να αναγνωρίσει το καπέλο αποφοίτου και τη φούντα στο κεφάλι του μικρού αδερφού του. Με μια άλλη λεζάντα πάγωσε. Το εξώφυλλο του άλμπουμ ήταν μια σέλφι του Αλσινέ και ενός φίλου του. Ο Αλσινέ έχει καταγωγή από την Αμερική και την Αϊτή· και ο φίλος του και ο ίδιος είναι μαύροι.

«Γορίλες», έλεγε.

«Έτσι σκέφτηκα — για να είμαι ειλικρινής, σκέφτηκα ότι κάτι είχα κάνει *εγώ*». Άνοιξε το άλμπουμ, περιμένοντας να έχει κάνει λάθος κλικ ή να έχει βάλει λάθος ετικέτα. Το άλμπουμ ήταν γεμάτο φωτογραφίες του Αλσινέ και του φίλου του. Και τίποτα άλλο. «Λέω — Είχε πάνω

από εβδομήντα φωτογραφίες. Δεν υπάρχει περίπτωση… Εκείνη τη στιγμή ακριβώς κατάλαβα τι είχε συμβεί».

Ο Αλσινέ μπήκε στο Twitter. «Φωτογραφίες Google», έγραψε, «τα κάνατε μαντάρα. Ο φίλος μου δεν είναι γορίλας».[23]

Μέσα σε δύο ώρες, ο κύριος αρχιτέκτονας της Google+, ο Γιόναταν Ζούνγκερ, επικοινώνησε μαζί του. «Ο Χριστός και η Παναγία», έγραψε. «Αυτό είναι 100% προβληματικό».

Η ομάδα του Ζούνγκερ εφάρμοσε μια αλλαγή στις Φωτογραφίες Google μέσα σε λίγες ώρες, και το επόμενο πρωινό μόνο δύο φωτογραφίες είχαν ακόμη λάθος ετικέτα. Έπειτα η Google έλαβε πιο δραστικά μέτρα: αφαίρεσε εντελώς την ετικέτα.

Για την ακρίβεια, τρία χρόνια αργότερα, το 2018, το περιοδικό Wired ανέφερε ότι η ετικέτα «γορίλας» ήταν ακόμη απενεργοποιημένη στις Φωτογραφίες Google. Αυτό σημαίνει ότι, χρόνια αργότερα, τίποτα δεν θα έχει την ετικέτα «γορίλας», συμπεριλαμβανομένων των γοριλών.[24]

Κατά περίεργο τρόπο, ο Τύπος το 2018, όπως ακριβώς και το 2015, φαινόταν επανειλημμένα να χαρακτηρίζει τη φύση του σφάλματος λανθασμένα. Τα πρωτοσέλιδα ανακοίνωναν, «Δύο χρόνια μετά, η Google λύνει το πρόβλημα του «ρατσιστικού αλγορίθμου» αφαιρώντας την ετικέτα "γορίλας" από το εργαλείο ταξινόμησης εικόνων»· «η Google "διόρθωσε" τον ρατσιστικό αλγόριθμο αφαιρώντας τους γορίλες από την τεχνολογία επισήμανσης εικόνων»· και «ο "ρατσιστικός αλγόριθμος" των Φωτογραφιών Google έχει διορθωθεί αλλά όχι με τον καλύτερο τρόπο».[25]

Ως προγραμματιστής ο ίδιος και εξοικειωμένος με τα συστήματα μηχανικής μάθησης, ο Αλσινέ γνώριζε ότι το πρόβλημα δεν ήταν ένας προκατειλημμένος *αλγόριθμος*. (Ο αλγόριθμος ήταν στοχαστική κατάβαση δυναμικού, η πιο γενική, συντηρητική, για όλες τις χρήσεις ιδέα στην επιστήμη των υπολογιστών: πέρνα τα δεδομένα εκπαίδευσης στην τύχη, ρύθμισε τις παραμέτρους του μοντέλου για να υπάρχουν ελαφρώς καλύτερες πιθανότητες για να πετύχεις τη σωστή κατηγορία για την εκάστοτε εικόνα και επανάλαβε όπως χρειάζεται.) Όχι, αυτό που κατάλαβε αμέσως ήταν ότι κάτι είχε πάει εντελώς στραβά στα ίδια τα δεδομένα εκπαίδευσης. «Δεν μπορούσα καν να κατηγορήσω τον αλγόριθμο», λέει. «Δεν φταίει ο αλγόριθμος. Έκανε ακριβώς αυτό για το οποίο σχεδιάστηκε».

Το πρόβλημα, φυσικά, σε ένα σύστημα που μπορεί, θεωρητικά, να μάθει τα πάντα από ένα σύνολο παραδειγμάτων είναι ότι έπειτα θα βρεθεί στο έλεος των παραδειγμάτων από τα οποία έμαθε.