

Τεχνητή Νοημοσύνη - Εισαγωγή

Καθ. Βασίλης Μεγαλοικονόμου

ΤΜΗΥΠ

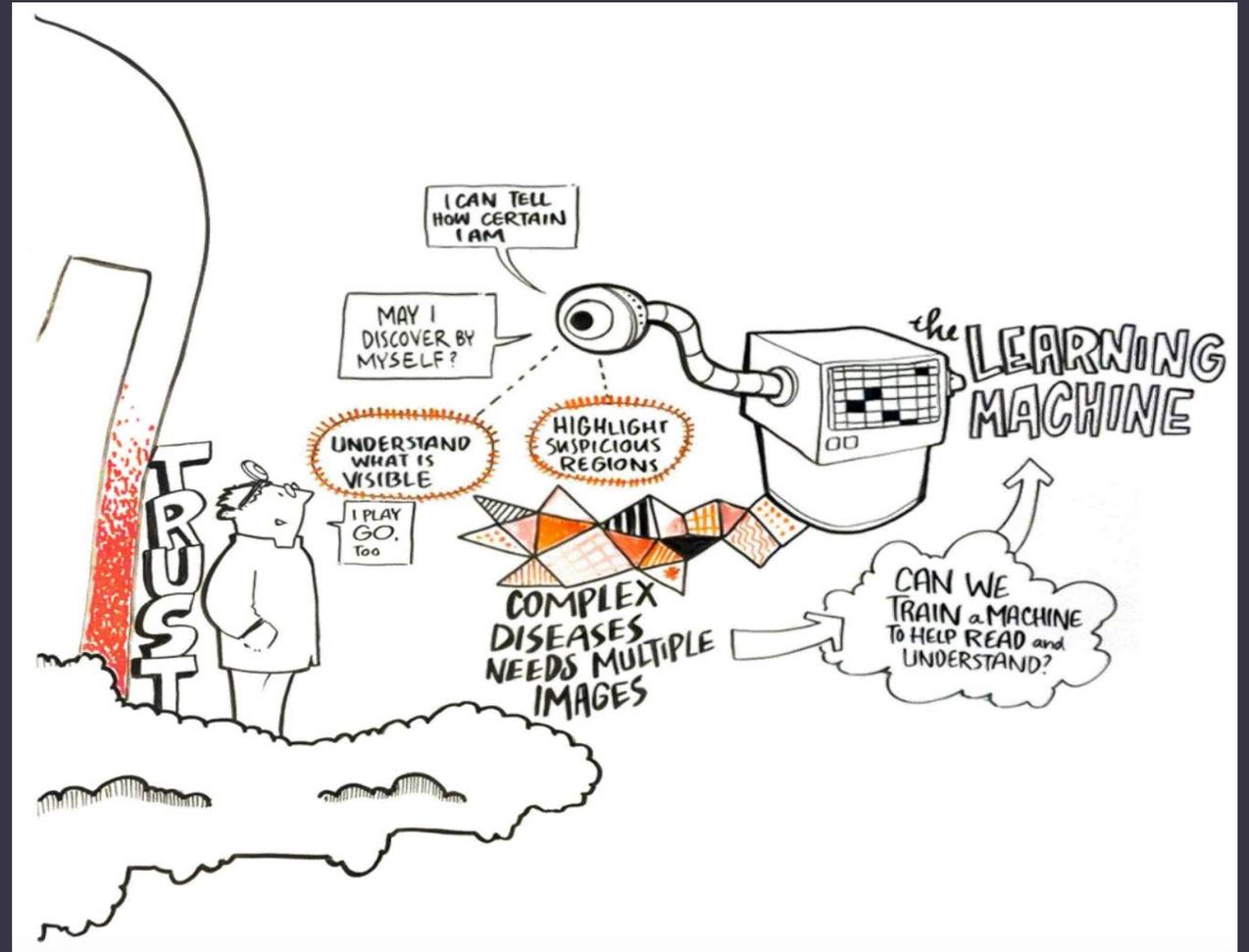
Παν. Πατρών



Ορισμός της Τεχνητής Νοημοσύνης

Δημιουργία λογικής και υλοποίησής της σε λογισμικό και υλικό των υπολογιστών με επιτυχία έτσι ώστε:

- Να μπορεί να μιμηθεί την ανθρώπινη νοημοσύνη (γνώση)
- Να αναπαράγει ανθρώπινη συμπεριφορά για μια συγκεκριμένη εργασία
- Να βρίσκει σχέσεις μεταξύ δεδομένων που συσχετίζονται ή έχουν αιτιώδη σχέση με το αποτέλεσμα



Κάποια ιστορικά στοιχεία

In (2000) 70% chances for an average human of not being able to distinguish human answers from computer generated answers in a conventional dialogue



Alan Turing

1950

1960

In (1970) computers exploring AI will be able to become world champions of chess



Allen Newell & Herbert Simon



Perfection of Personal Assistants

2025?

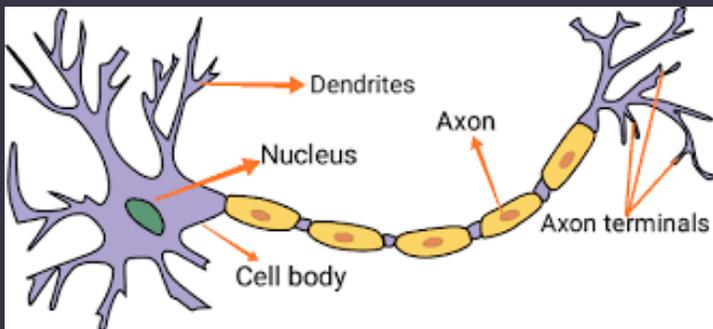
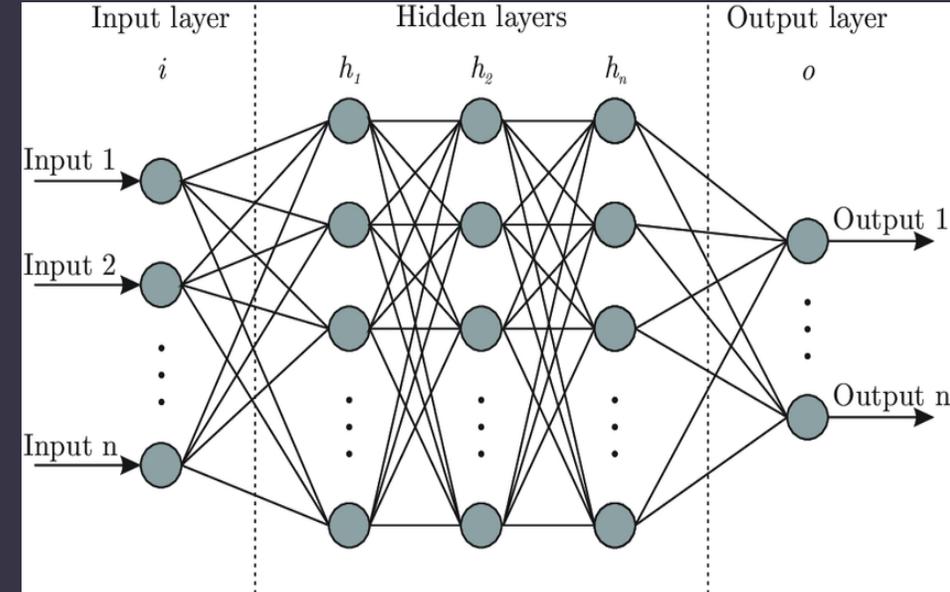
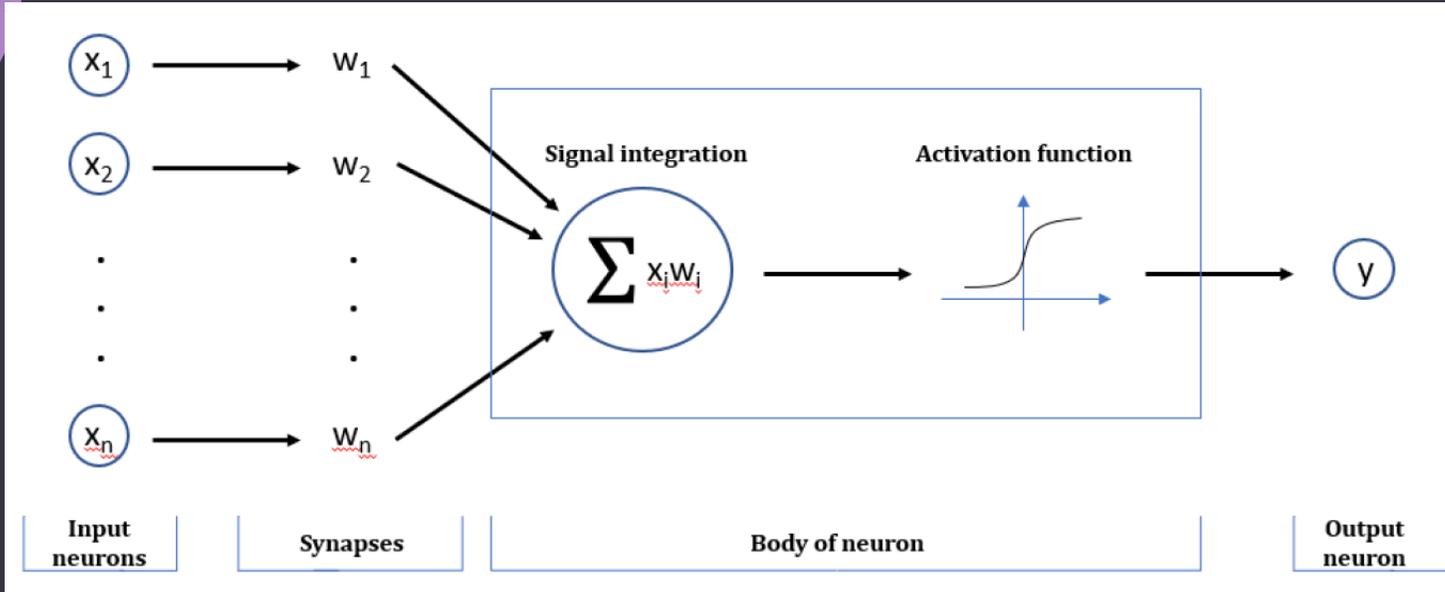
1997



2015



Πως μπορούμε να φτιάξουμε έξυπνα συστήματα;

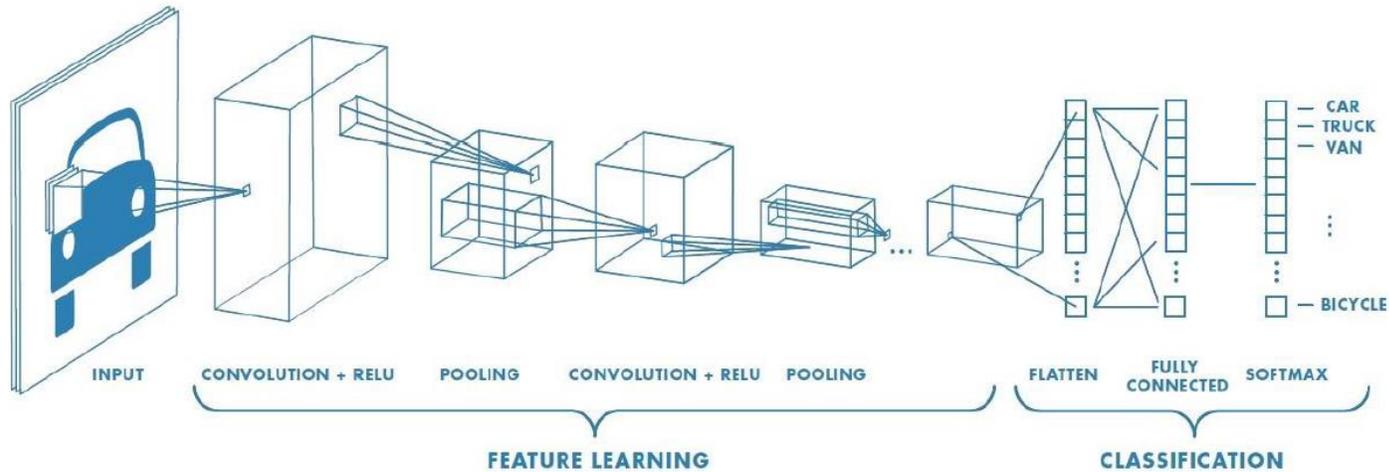


Discriminative Models (data first):

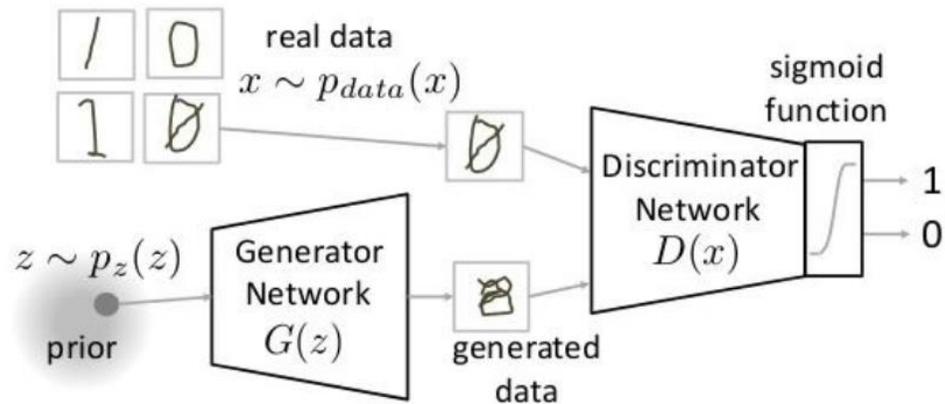
- πολύ περίπλοκα μοντέλα που δεν ακολουθούν την ανθρώπινη διαίσθηση
- δεν μπορούν να ερμηνευθούν
- αναπαράγουν όσο μπορούν τις παρατηρήσεις

Βαθιά Νευρωνικά Δίκτυα: Διάφορες αρχιτεκτονικές

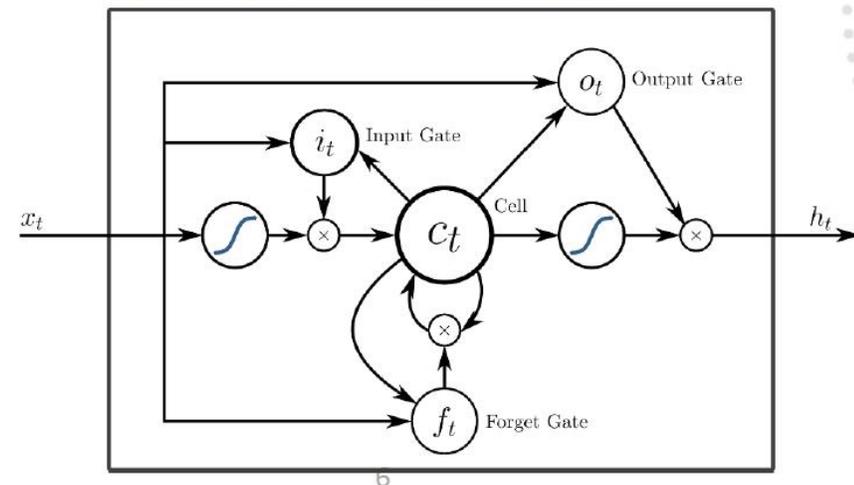
Convolutional Neural Networks



Generative (& Adversarial) Neural Networks

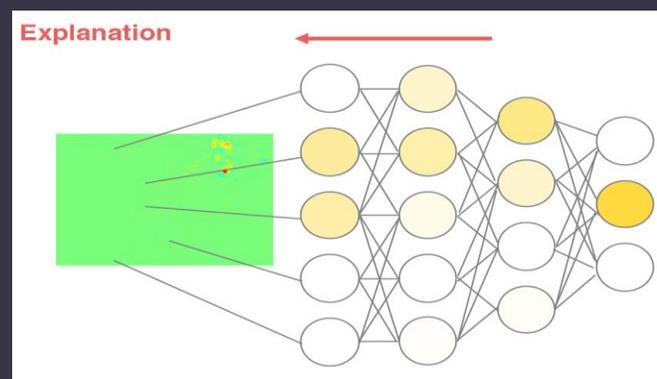
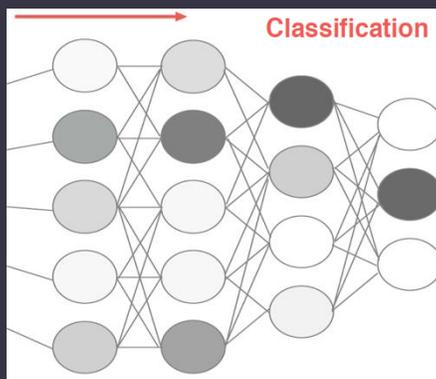
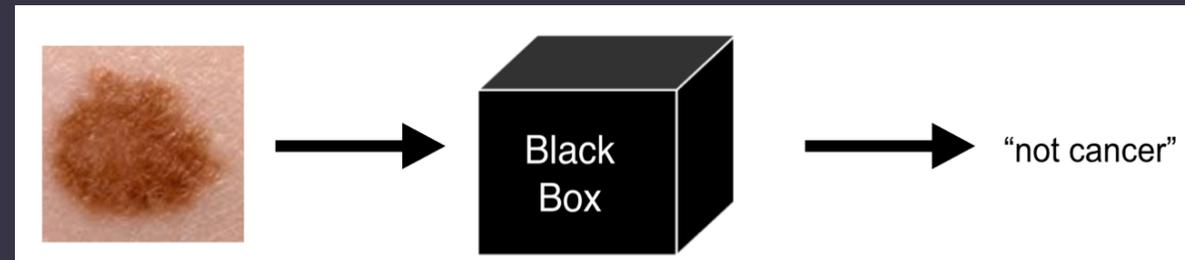


Long Short Memory Networks



Τα μοντέλα TN είναι μαύρα κουτιά;

- Εξηγώντας τις προβλέψεις:
 - Θέλουμε να μπορούμε να εξηγήσουμε γιατί ένα συγκεκριμένο μοτίβο έχει ταξινομηθεί με συγκεκριμένο τρόπο
 - Ποιες διαστάσεις των δεδομένων είναι πιο σχετικές για την συγκεκριμένη εργασία;
 - Ποια έννοια κωδικοποιεί ένας συγκεκριμένος νευρώνας;
- Δεν είναι όλα τα μοντέλα μηχανικής μάθησης μαύρα κουτιά! Π.χ. δέντρα αποφάσεων, Bayesian δίκτυα, κ.α.
- Διάδοση συνάφειας κατά επίπεδο:
- Ιδέα: Αναδιανείμετε τα στοιχεία για την κατηγορία πίσω στον αρχικό χώρο (εικόνα).

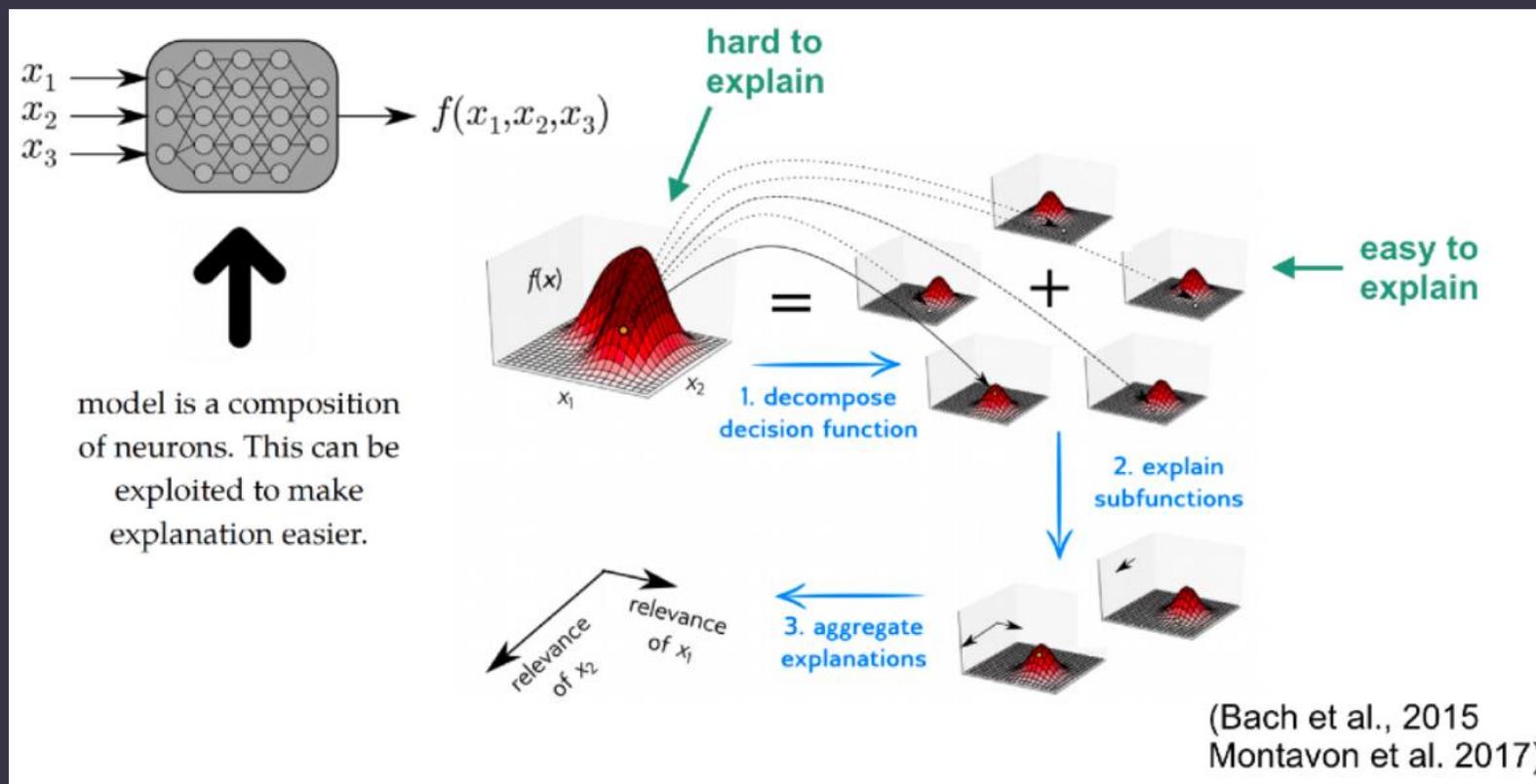


Τα μοντέλα TN είναι μαύρα κουτιά;

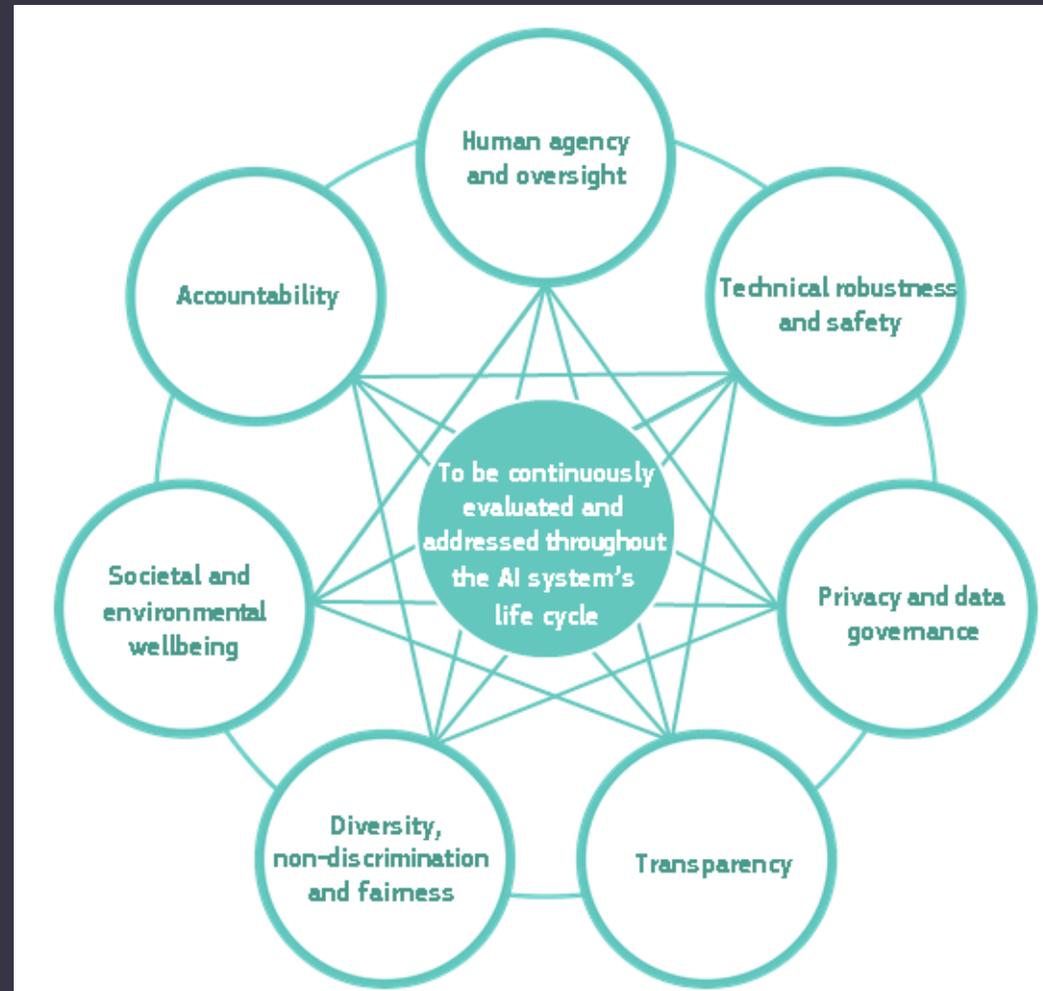
Διάδοση συνάφειας κατά επίπεδο (Layer-wise relevance propagation -LRP):

Από τις πρώτες εργασίες σε ερμηνευσιμότητα των Deep NNs.

Για να εξηγήσει με ασφάλεια ένα μοντέλο αξιοποιεί τη δομή του νευρικού δικτύου της συνάρτησης απόφασης



Αξιόπιστη ΤΝ: απαιτήσεις κλειδιά για συστήματα ΤΝ



High-Level Expert Group on AI set up by the European Commission



Αξιόπιστη ΤΝ

Η ομάδα εμπειρογνομένων υψηλού επιπέδου για την τεχνητή νοημοσύνη (AI HLEG), μια ομάδα εμπειρογνομένων που διορίστηκε από την Ευρωπαϊκή Επιτροπή για να παρέχει συμβουλές σχετικά με τη στρατηγική της για τεχνητή νοημοσύνη, κυκλοφόρησε πρόσφατα ένα έγγραφο με οδηγίες για την επίτευξη «αξιόπιστης τεχνητής νοημοσύνης» που αναφέρει επτά βασικές απαιτήσεις, με διαφάνεια είναι μία από αυτές τις απαιτήσεις.

Αξιόπιστη ΤΝ: απαιτήσεις κλειδιά για συστήματα ΤΝ

Ανθρώπινη παρέμβαση και εποπτεία

Τεχνική στιβαρότητα και ασφάλεια: Ανθεκτικότητα σε επιθέσεις, ακρίβεια, αξιοπιστία, αναπαραγωγιμότητα

Ιδιωτική ζωή και διακυβέρνηση των δεδομένων: Προστασία ιδιωτικής ζωής και δεδομένων, ποιότητα και ακεραιότητα δεδομένων, πρόσβαση στα δεδομένα

Διαφάνεια: Ιχνηλασιμότητα, επεξηγησιμότητα, επικοινωνία

Πολυμορφία, δικαιοσύνη και απαγόρευση διακρίσεων: αποφυγή αθέμιτης μεροληψίας, προσβασιμότητα, καθολικός σχεδιασμός

Κοινωνική και περιβαλλοντική ευημερία: κοινωνικές επιπτώσεις, ΤΝ βιώσιμη και φιλική προς το περιβάλλον

Λογοδοσία: Ελεγχιμότητα, ελαχιστοποίηση και γνωστοποίηση αρνητικών επιπτώσεων, αντισταθμιστικές ρυθμίσεις, έννομη προστασία



Αξιόπιστη ΤΝ

FUTURE-AI Guiding Principles



Fairness

for equitable



Universality

for standardised



Traceability

for monitoring



Usability

for transferable



Robustness

for reliable



Explainability

for interpretable

AI

SOLUTIONS
in
MEDICAL IMAGING

Δικαιοσύνη για Equitable AI

- Ανισότητες μεταξύ ατόμων/ομάδων ατόμων, λόγω διαφορών σε φύλο, ηλικία, εθνικότητα, εισόδημα, εκπαίδευση, γεωγραφία.
- Σύνολα δεδομένων εκπαίδευσης: ποσοτική και ποιοτική ποικιλομορφία και ισορροπία - ίδια απόδοση σε υποπληθυσμούς.
- Ζητήματα δικαιοσύνης στην ενσωμάτωση της TN στην καθημερινή πρακτική
 - χρήση από έμπειρους ή λιγότερο έμπειρους χρήστες και επίδραση στις ικανότητές τους στη λήψη αποφάσεων
- Εντοπισμός κρίσιμων βημάτων ή υποσυστημάτων που απαιτούν λήψη αποφάσεων με «**human in the loop**» και ανατροφοδότηση από τους χρήστες - αποφυγή μεροληψίας από τον αυτοματισμό
- Μηχανισμοί όπως η τυχαία δειγματοληψία, η στρωματοποιημένη δειγματοληψία και η προσαρμοστική δειγματοληψία - αύξηση της ισορροπίας και της αντιπροσωπευτικότητας των δεδομένων.
- Πολυκεντρικά σύνολα δεδομένων εκπαίδευσης/δοκιμής
- Διαφάνεια δικαιοσύνης
 - διαφανής και τεκμηριωμένη διαδικασία συλλογής/προετοιμασίας των συνόλων δεδομένων για εκπαίδευση/δοκιμή λύσεων TN, συμπεριλαμβανομένων πληροφοριών σχετικά με ποικιλομορφία και ισορροπία δεδομένων.
- Συνεχής παρακολούθηση της δικαιοσύνης: Κατά την εφαρμογή ο αλγόριθμος TN θα πρέπει να αξιολογείται διεξοδικά και συνεχώς και να επανεκπαιδεύεται για δικαιοσύνη.
- Εκπαιδευτικό υλικό για στοχευμένους τελικούς χρήστες λύσεων TN

Καθολικότητα για τυποποιημένη ΤΝ

- Ορισμός και εφαρμογή **προτύπων** κατά την ανάπτυξη, αξιολόγηση και χρήση των αλγορίθμων
 - τα πρότυπα, συμπεριλαμβάνουν τα τεχνικά, ειδικά πρότυπα τομέα εφαρμογής, δεοντολογικά και κανονιστικά πρότυπα
- Τυποποίηση: εμφανή οφέλη για τη διαλειτουργικότητα, υιοθέτηση και εμπιστοσύνη
- Τυποποίηση λογισμικού: τα πλαίσια βοηθούν στην αποφυγή πιθανών ζητημάτων ασυμβατότητας
- Η τυποποίηση στις αναφορές, την επισήμανση και τον σχολιασμό διασφαλίζει την πληρότητα και τη διαλειτουργικότητα των συνόλων δεδομένων
- Χρήση τυπικών κριτηρίων και μετρήσεων για τις αξιολογήσεις
- Σύνολα δεδομένων αναφοράς για συγκριτική αξιολόγηση μεθόδων ΤΝ
- Οι βασικές λεπτομέρειες των αλγορίθμων αναφέρονται σαφώς

Ιχνηλασιμότητα για διαφανή TN

- Τεκμηρίωση όλης της διαδικασίας ανάπτυξης και παρακολούθηση της λειτουργίας ενός μοντέλου/συστήματος TN
- Η «διαφάνεια» του μοντέλου και η “ιχνηλασιμότητα βάσει σχεδίου” - κλειδί για αποφυγή οποιασδήποτε «γκρίζας» περιοχής ... σχετικά με το τι συμβαίνει εάν κάτι πάει λάθος όταν το μοντέλο χρησιμοποιείται στην πράξη
- Η διαφάνεια στην ανάπτυξη και χρήση TN απαιτεί σαφή επικοινωνία μιας ποικιλίας εργασιών:
 - διαχείριση δεδομένων
 - ανάπτυξη και ενημέρωση/βελτίωση μοντέλων
 - εργασίες που σχετίζονται με τις λειτουργικές λεπτομέρειες του συστήματος
- Μεγάλη σημασία στην προέλευση των δεδομένων και στην παρακολούθηση ολόκληρου του κύκλου ζωής του μοντέλου
- **Διαφάνεια δεδομένων** : η διαφάνεια στη συλλογή, χρήση και αποθήκευση δεδομένων
 - Μέθοδοι προέλευσης δεδομένων (ή γενεαλογίας δεδομένων) για τη βελτίωση της αναπαραγωγής, της ανίχνευσης, της αξιολόγησης ποιότητας στη χρήση δεδομένων και των διαδικασιών μετασχηματισμού δεδομένων

Ιχνηλασιμότητα για διαφανή TN (2)

- **Διαφάνεια μοντέλου** : π.χ. ModelOps , πλαίσιο βασισμένο στο cloud για διαχείριση pipelines TN από άκρο σε άκρο. Περιλαμβάνει:
 - σύνολα δεδομένων, ορισμούς μοντέλων, εκπαιδευμένα μοντέλα, εφαρμογές και συμβάντα παρακολούθησης, αλγόριθμους και πλατφόρμες για επεξεργασία δεδομένων, εκπαίδευση μοντέλων ή ανάπτυξη εφαρμογών
- Η διαφάνεια και η ιχνηλασιμότητα σημαντικά για: αναπαραγωγιμότητα, δυνατότητα ελέγχου και τη λογοδοσία
- Συνεχής επιτήρηση μοντέλων TN και σύστημα συντήρησης
 - Παρακολούθηση απόδοσης, συμπεριφοράς στο χρόνο, ζωτικότητας, συμπεριφοράς των μοντέλων σε πραγματικές συνθήκες, απόκλιση από ρυθμίσεις εκπαίδευσης ή τις προηγούμενες καταστάσεις
- Βρόχος ανάδρασης δεδομένων
 - αξιοποίηση νέων δεδομένων, νέας γνώσης και ανατροφοδότηση από πραγματικές ρυθμίσεις παραγωγής
 - Απόδοση μπορεί να υποβαθμίζεται με τον χρόνο όταν αξιολογείται στον πραγματικό κόσμο
- Ιχνηλασιμότητα μέσω πλαισίου διακυβέρνησης για ολόκληρο τον κύκλο ζωής του μοντέλου

Ιχνηλασιμότητα για διαφανή TN (3)

Συστάσεις για ιχνηλασιμότητα :

- Πεδίο εφαρμογής μοντέλου: όροι προβλεπόμενης χρήσης του μοντέλου, σενάρια/περιπτώσεις χρήσης, επιδιωκόμενο αποτέλεσμα, υποστηριζόμενα inputs των μοντέλων, τυχόν γνωστοί περιορισμοί του προβλήματος που αντιμετωπίζει
- Προέλευση δεδομένων: συμπεριλαμβανομένων πληροφοριών σχετικά με την προέλευση και την ιδιοκτησία των δεδομένων, τα πρωτόκολλα απόκτησης, τις συσκευές και το timing
- Παρακολούθηση της θέσης των δεδομένων στο δίκτυο
- Τεκμηρίωση της προετοιμασίας των δεδομένων
- Καταγραφή εκπαίδευσης του μοντέλου
- Τεκμηρίωση επικύρωσης: Η διαδικασία επικύρωσης θα πρέπει να περιγράφεται δεόντως ως προς τις μετρικές αξιολόγησης, την προσέγγιση διασταυρούμενης επικύρωσης, κ.λπ.
- Εργαλεία ιχνηλασιμότητας:
 - επιτρέπουν την παρακολούθηση της ζωντανής λειτουργίας του εργαλείου, την επισήμανση/καταγραφή σφαλμάτων, αποκλίσεων, υποβάθμισης στην απόδοση, εξέλιξης του μοντέλου με την πάροδο του χρόνου.
- Διαβατήριο μοντέλων

Ευχρηστία για αποτελεσματική/ευεργετική ΤΝ

- «Ο βαθμός στον οποίο ένα προϊόν μπορεί να χρησιμοποιηθεί από συγκεκριμένους χρήστες για την επίτευξη συγκεκριμένων στόχων με αποτελεσματικότητα, αποδοτικότητα και ικανοποίηση σε ένα συγκεκριμένο πλαίσιο χρήσης».
- Σχεδίαση με **επίκεντρο τον χρήστη** για καθεμία από τις φάσεις του κύκλου ζωής του μοντέλου
- Η χρηστικότητα συνδέεται παραδοσιακά με διάφορα χαρακτηριστικά:
 - Δυνατότητα εκμάθησης: πόσο γρήγορα ένας νέος χρήστης μαθαίνει την χρήση - κρίσιμο για γρήγορη υιοθέτηση
 - Αποδοτικότητα: μειώνοντας τον βαρύ φόρτο εργασίας, παρέχοντας παραγωγικότητα
 - Δυνατότητα εύκολης μνημόνευσης
 - Περιορισμένα και μη καταστροφικά λάθη
 - Ικανοποίηση: (υποκειμενικό) παρακολούθηση για διασφάλιση της ευρείας υιοθέτησης
- Συμμετοχή ειδικών τομέα στο σχεδιασμό - διαφορετικά υπάρχει ο κίνδυνος να γίνουν άσχετες υποθέσεις
- Ενεργή εμπλοκή διεπιστημονικών ομάδων
- Κατανόηση αναγκών των χρηστών - Σχεδιασμός διεπαφής χρήστη
- Επεξήγηση για χρηστικότητα. Δοκιμή χρηστικότητας. Συνεχής παρακολούθηση της ικανοποίησης των χρηστών
- Παροχή εκπαιδευτικών πόρων για τελικούς χρήστες

Στιβαρότητα για αξιόπιστη ΤΝ

- Ικανότητα μιας τεχνολογίας ΤΝ να διατηρεί την ακρίβεια του μοντέλου της όταν εφαρμόζεται υπό εξαιρετικά μεταβλητές συνθήκες στον πραγματικό κόσμο, έξω από το ελεγχόμενο περιβάλλον του εργαστηρίου όπου είναι κατασκευασμένος ο αλγόριθμος
- Ικανότητα να αντιμετωπίζει προβλήματα ετερογένειας δεδομένων
- Συστάσεις για ανθεκτικότητα:
 - Εναρμόνιση δεδομένων : Εάν οι διαφορές στα πρωτόκολλα απόκτησης δεδομένων δεν μπορούν να αποφευχθούν μεταξύ των κέντρων
 - Ποιοτικός έλεγχος: θα πρέπει να εφαρμόζεται για τον εντοπισμό μη φυσιολογικών αποκλίσεων ή τεχνουργημάτων
 - Αύξηση δεδομένων για εκπαίδευση μοντέλων
 - μέσω συνθετικών δεδομένων με προσομοίωση ενός ευρέος φάσματος δύσκολων συνθηκών
 - Εκπαίδευση σε ετερογενή δεδομένα
 - Εκτίμηση αβεβαιότητας

Επεξηγησιμότητα για βελτιωμένη κατανόηση της ΤΝ

Επεξηγησιμότητα – Ερμηνευσιμότητα:

- Οι λύσεις ΤΝ γενικά, και τα βαθιά νευρωνικά δίκτυα ειδικότερα, στερούνται διαφάνειας
- «Μαύρο κουτί ΤΝ»: τα μοντέλα μαθαίνουν πολύπλοκες λειτουργίες που είναι απρόσιτες και συχνά ακατανόητες για τον άνθρωπο
- GDPR: Ο Γενικός Κανονισμός για την Προστασία Δεδομένων της Ευρωπαϊκής Ένωσης καθορίζει το **«δικαίωμα στην εξήγηση»**
- Η εξηγήσιμη ΤΝ (ΧΑΙ) αναφέρεται σε λύσεις ΤΝ που δίνουν στους τελικούς χρήστες πληροφορίες για τη λειτουργία της
 - η επεξήγηση των λύσεων ΤΝ ξεκινά από τη διαδικασία σχεδιασμού και συγκέντρωσης απαιτήσεων
 - ενσωματώνει τις επιθυμίες, τους στόχους και τις προκλήσεις των τελικών χρηστών για να κατανοήσουν ποιος τύπος επεξηγήσεων ταιριάζει καλύτερα στις ανάγκες τους