Ηθική και Τεχνητή Νοημοσύνη

Κάτια Κερμανίδου

kerman@ionio.gr

Τι είναι Ηθική (Ethics);

- Προέρχεται από την αρχαία ελληνική λέξη «**ήθος**», που σημαίνει χαρακτήρας, προσωπικότητα, συνήθεια, τρόπος ζωής
- **Ηθική** είναι η κατανόηση και η συμφωνία των ανθρώπων σχετικά
  - με το τι είναι σωστό και τι είναι λάθος, καλό ή κακό
  - με τους κανόνες που έχουν τεθεί και πρέπει οι άνθρωποι και η κοινωνία να ακολουθούν
- Η Ηθική στην ΤΝ αναφέρεται στις ηθικές αρχές και αξίες που πρέπει να διέπουν την σχεδίαση, ανάπτυξη και χρήση συστημάτων ΤΝ
- Βασικές αρχές Ηθικής αποτελούν
  - Η δικαιοσύνη
  - Η διαφάνεια
  - Η ιδιωτικότητα
  - Η ασφάλεια
  - Η ισότητα
- Τα συστήματα ΤΝ πρέπει να σχεδιάζονται και να χρησιμοποιούνται με τρόπο τέτοιο ώστε
  - να επωφελείται η κοινωνία
  - να ακολουθούνται οι αρχές ηθικής

Γιατί έχει σημασία η Ηθική στην ΤΝ;

Colombian judge says he used ChatGPT in ruling

Juan Manuel Padilla asked the AI tool how laws applied in case of autistic boy's medical funding, while also using precedent to support his decision

SLAUGHTERBOTS ARE HERE.

The era in which algorithms decide who lives and who dies is upon us. We must act now to prohibit and regulate these weapons.

sky news

'Regulate it before we're all finished': Musicians react to AI songs flooding the internet

Η οικογένεια του πρώην άσου της F1 Μίκαελ Σουμάχερ μηνύει γερμανικό περιοδικό που δημοσίευσε ψεύτικη συνέντευξή του με τη βοήθεια της τεχνητής νοημοσύνης (AI). Είναι το τελευταίο κρούσμα που φωτίζει τα ηθικά και νομικά όρια της AI. Φωτ. EPA / HANNIBAL HANSCHKE

Deloitte to refund government, admits using AI in $440k report

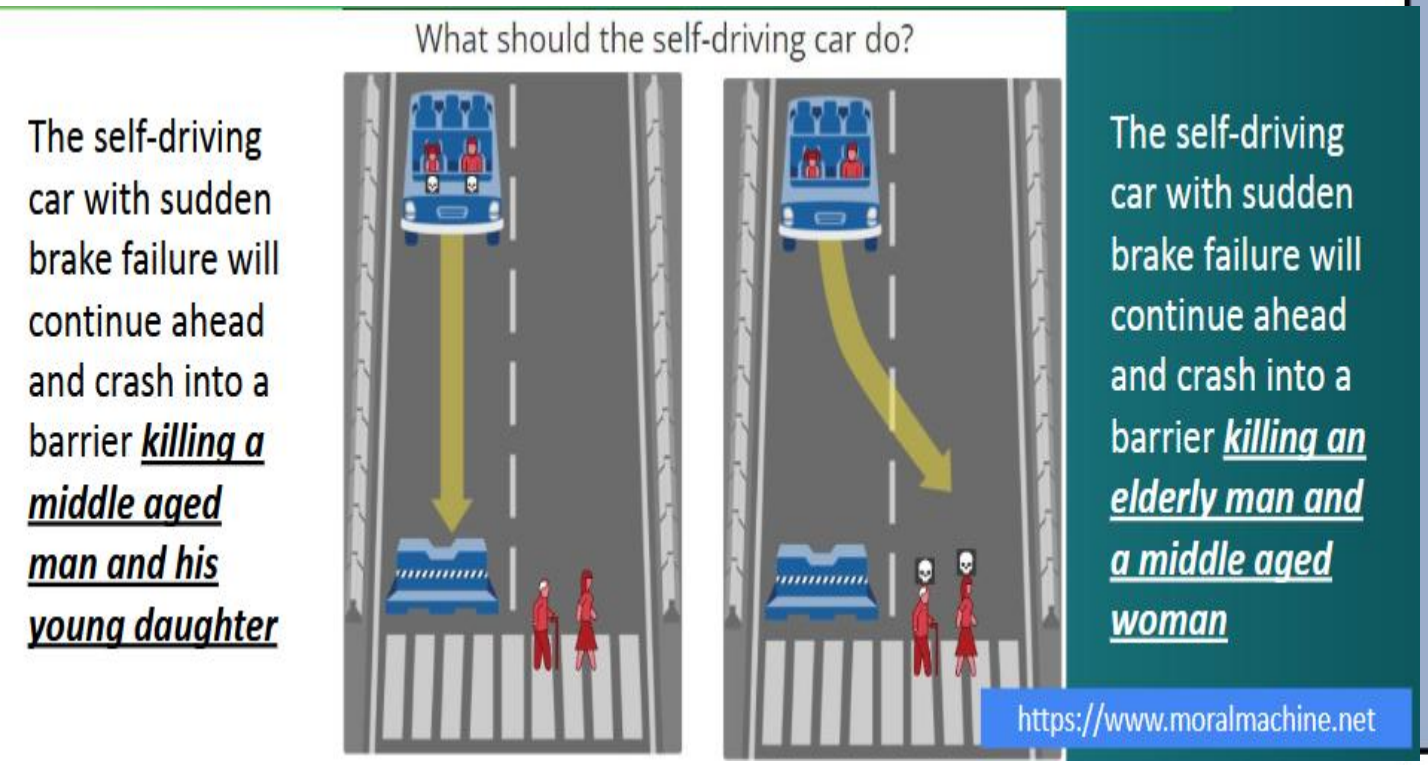
Edmund Tadros and Paul Karp

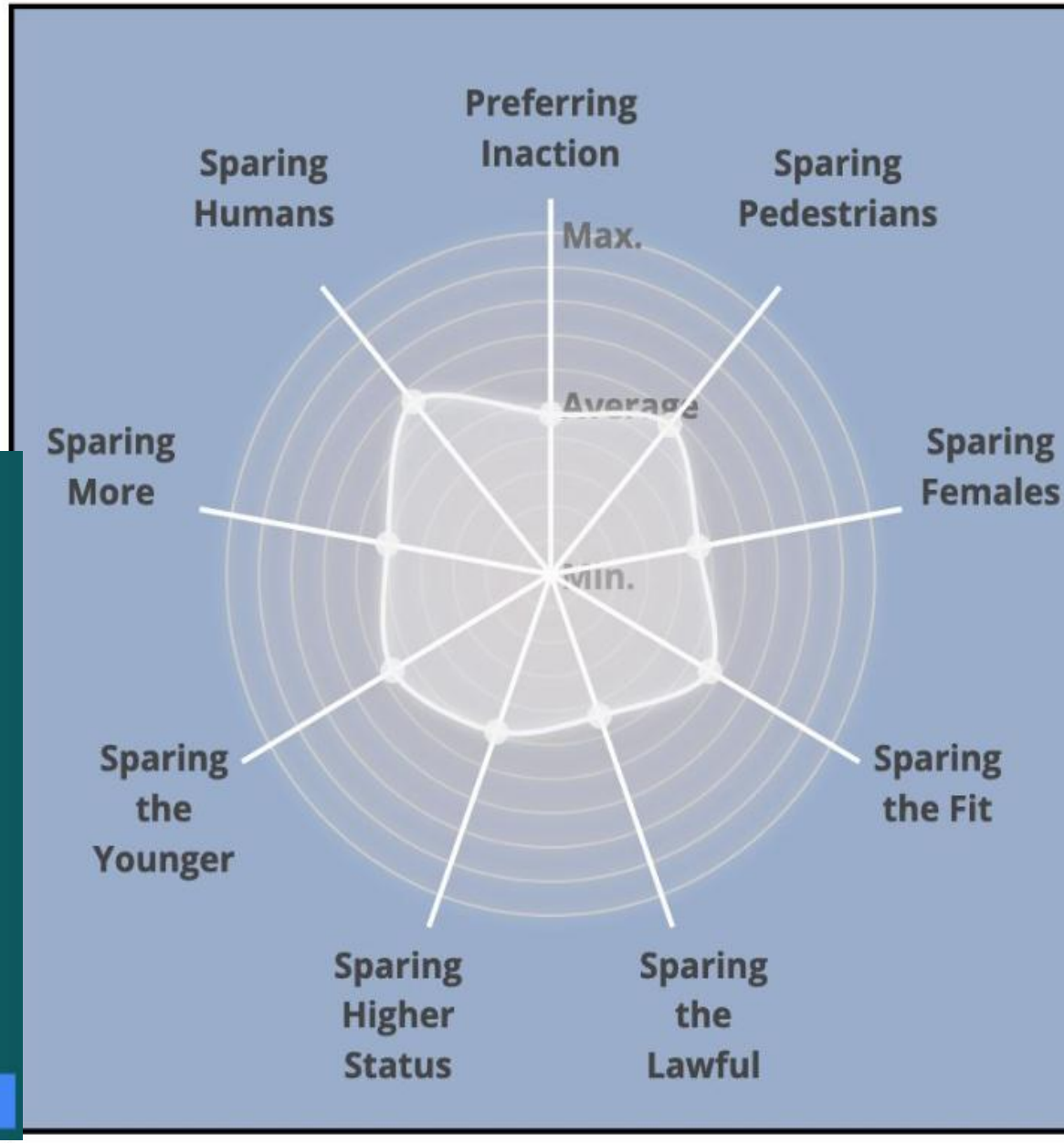
Oct 5, 2025 – 7.41pm

Deloitte Australia will issue a partial refund to the federal government after admitting that artificial intelligence had been used in the creation of a $440,000 report littered with errors including three nonexistent academic references and a made-up quote from a Federal Court judgement.

A new version of the report for the Department of Employment and Workplace Relations (DEWR) was quietly uploaded to the department's website on Friday, ahead of a long weekend across much of Australia. It features more than a dozen deletions of nonexistent references and footnotes, a rewritten reference list, and corrections to multiple typographic

Γιατί έχει σημασία η Ηθική στην ΤΝ;





The grey area is the world average

https://www.adaptcentre.ie/wp-content/uploads/2024/01/AI-and-Ethics.pdf

Γιατί έχει σημασία η Ηθική στην ΤΝ;

- Σε ένα σύστημα διαχείρισης της λίστας ασθενών για μεταμόσχευση οργάνου
- Ποιος έχει προτεραιότητα;
- Κάποιος διάσημος/ζάμπλουτος;
  - Θέματα Ηθικής: ιδιωτικότητα, εμπιστευτικότητα, εκβιασμός, εμπορική εκμετάλλευση
- Αριστοτέλης
  - Δικαισύνη σημαίνει κάθε άνθρωπος να παίρνει αυτό που (του) αξίζει
    - δεδομένου του κατά πόσο υπακούει στους νόμους οι οποίοι είναι φτιαγμένοι για το κοινό καλό
    - δεδομένου του κατά πόσο προσφέρει στα κοινά
    - δεδομένου του κατά πόσο ζει ενάρετα

https://www.adaptcentre.ie/wp-content/uploads/2024/01/AI-and-Ethics.pdf

Γιατί έχει σημασία η Ηθική στην ΤΝ;

- Η ΤΝ σήμερα στηρίζεται σε **δεδομένα**
- Τα δεδομένα μπορεί να είναι
  - πολωμένα (biased)
    - Ένα σύστημα ΤΝ που εκπαιδεύεται με βιογραφικά των τελευταίων 10 ετών για μια θέση μηχανικού, όπου το 90% των προσληφθέντων ήταν άνδρες, θα «μάθει» να βαθμολογεί χαμηλότερα τα βιογραφικά γυναικών
    - Ένα μοντέλο που χρησιμοποιεί ταχυδρομικούς κώδικες για να εγκρίνει δάνεια, αποκλείοντας περιοχές με χαμηλότερο εισόδημα, κάνει διακρίσεις κατά συγκεκριμένων κοινωνικών ομάδων
    - Αν ένα μοντέλο ανάλυσης συναισθήματος εκπαιδευτεί μόνο με tweets από μία πολιτική παράταξη, θα παρερμηνεύει τη θετική ή αρνητική χροιά των δημοσιεύσεων της αντίθετης παράταξης
    - Αν μια κλινική μελέτη για ένα νέο φάρμακο περιλαμβάνει μόνο ασθενείς 50-60 ετών, τα αποτελέσματα (δεδομένα) δεν θα είναι αντιπροσωπευτικά για τη δράση του φαρμάκου σε νεότερους ή γηραιότερους ανθρώπους
  - ψευδή (false)
    - Ψευδείς ειδήσεις
    - Φάρσες/ψεύτικα email

Γιατί έχει σημασία η Ηθική στην ΤΝ;

- Η ΤΝ σήμερα εξαπλώνεται όλο και περισσότερο, και ο αντίκτυπός της στην κοινωνία, στην οικονομία, στην εκπαίδευση είναι όλο και μεγαλύτερος
- Στην ΤΝ στηρίζεται σήμερα όλο και περισσότερο η λήψη σοβαρών αποφάσεων που επηρεάζουν την ζωή των ανθρώπων και την κοινωνία
  - Αποκτά όλο και περισσότερη **αυτονομία**
- Όσο η αυτονομία των εργαλείων ΤΝ μεγαλώνει, τόσο μεγαλύτερη είναι η ανάγκη
  - για περισσότερο και καλύτερο **ανθρώπινο έλεγχο** των εργαλείων
  - για πιο καλή **ερμηνευσιμότητα (explainability)** των αποτελεσμάτων της ΤΝ
    - γιατί είναι αυτά τα αποτελέσματα;
  - για πιο διάφανη **ικανότητα λογοδοσίας** (**accountability**) στα αποτελέσματα της ΤΝ
    - ποιος ευθύνεται για αυτά τα αποτελέσματα;

Αρχές Ηθικής για την Ανάπτυξη Συστημάτων ΤΝ

- Ανθρώπινη Επίβλεψη (Human Oversight)
  - Όλες οι αποφάσεις αξιολογούνται από ανθρώπους και άνθρωπος είναι υπεύθυνος για τις τελικές αποφάσεις
- Τεχνική Ευρωστία (Technical Robustness)
  - Τα εργαλεία ΤΝ δεν περιλαμβάνουν σφάλματα κατά τη σχεδίασή τους
- Ασφάλεια (Safety)
  - Η λειτουργία των εργαλείων ΤΝ δεν οδηγεί σε κίνδυνο, δεν βλάπτει
- Ιδιωτικότητα (Privacy)
  - Αυστηροί κανόνες διέπουν τα ανθρώπινα δεδομένα, και τον τρόπο με τον οποίο τα εργαλεία ΤΝ τα χρησιμοποιούν
- Διαφορετικότητα (Diversity), Μη Διάκριση (Non-discrimination), Δικαιοσύνη (Fairness)
  - Η ΤΝ πρέπει να δουλεύει με τον ίδιο τρόπο για διαφορετικούς ανθρώπους και για διαφορετικά πλαίσια συμφραζομένων (contexts).
  - Τα πλεονεκτήματα της ΤΝ πρέπει να διαμοιράζονται με ισότητα

Αρχές Ηθικής για την Ανάπτυξη Συστημάτων ΤΝ (συν)

- Κοινωνική ευεξία (Societal Well Being)
  - Ο αντίκτυπος της ΤΝ στην κοινωνία πρέπει να είναι θετικός
- Περιβαλλοντική ευεξία (Environmental Well Being)
  - Ο αντίκτυπος της ΤΝ στο περιβάλλον πρέπει να είναι θετικός
- Διαφάνεια (Transparency)
  - Να έχουμε ξεκάθαρη επίγνωση τι αποφάσεις θα παίρνει η ΤΝ, ποια βήματα οδηγούν σε αυτές τις αποφάσεις, από τι δεδομένα προκύπτει η πληροφορία που τις πυροδοτεί
- Απόδοση Ευθύνης (Accountability)
  - Ξέρουμε ποιος είναι υπεύθυνος για τα αποτελέσματα της ΤΝ, μπορούμε να διορθώσουμε σφάλματα και να επιβάλουμε ποινή/τιμωρία όταν κάτι είναι ανήθικο/παράνομο.

Σενάριο: Ζητάω βοήθεια από το ChatGPT για την εργασία μου

- Πρέπει να ζητάω βοήθεια από την ΤΝ για την εργασία μου;
- Είναι το ίδιο με το να ζητάω βοήθεια από κάποιον φίλο/συμφοιτητή μου;
- Μπορεί να εδραιωθεί σχέση εμπιστοσύνης ανάμεσα σε φοιτητές και καθηγητές με αυτό τον τρόπο;

ΤΝ και Θέσεις Εργασίας

- Ο αυτοματισμός πάντα θεωρούταν απειλή για τις ανθρώπινες θέσεις εργασίας
  - Μηχανές, αυτόματες συσκευές, υπολογιστές αντικαθιστούν τους ανθρώπους εδώ και αιώνες
- Ρομπότ
  - αντικαθιστούν εργάτες σε εργοστάσια, αγρότες
- Εργαλεία ΤΝ
  - παράγουν κείμενο, βίντεο, εικόνες, μουσική
  - μεταφράζουν κείμενο
  - παράγουν κώδικα
  - παράγουν ιατρικές διαγνώσεις
  - διδάσκουν
  - προτείνουν νομικές αποφάσεις
- **Πώς εξασφαλίζουμε ότι επαγγέλματα δεν θα εξαφανιστούν;**
- **Πώς θα έχουμε όλοι ίσες ευκαιρίες σε αξιοπρεπή επαγγέλματα που εξασφαλίζουν ένα αποδεκτό βιωτικό επίπεδο;**

ΤΝ και Περιβάλλον

- Τα σύγχρονα εργαλεία ΤΝ απαιτούν μεγάλα **data centres** για την επεξεργασία των δεδομένων
  - κάθε τέτοιο κέντρο περιλαμβάνει πολλές χιλιάδες ή και εκατομμύρια από servers και υπολογιστές
- Αυτά τα κέντρα
  - είναι υπεύθυνα για πολύ μεγάλο ποσοστό της ενεργειακής κατανάλωσης σήμερα (>20%)
- Κάθε 5-50 prompts που απαντάει το chatgpt, καταναλώνει μισό λίτρο νερού για ψύξη των servers.
- Η πρώτη ύλη (μέταλλα και ορυκτά) που χρειάζονται για την κατασκευή των υπολογιστικών μονάδων βρίσκονται κυρίως σε αναπτυσσόμενες χώρες
  - η εξόρυξή τους είναι επιβλαβής για το περιβάλλον
  - και επικίνδυνη για τον τοπικό πληθυσμό

EU AI Act

An official website of the European Union How do you know?

European Commission

Shaping Europe's digital future

Home | Policies | Activities | News | Library | Funding | Calendar | Consultations

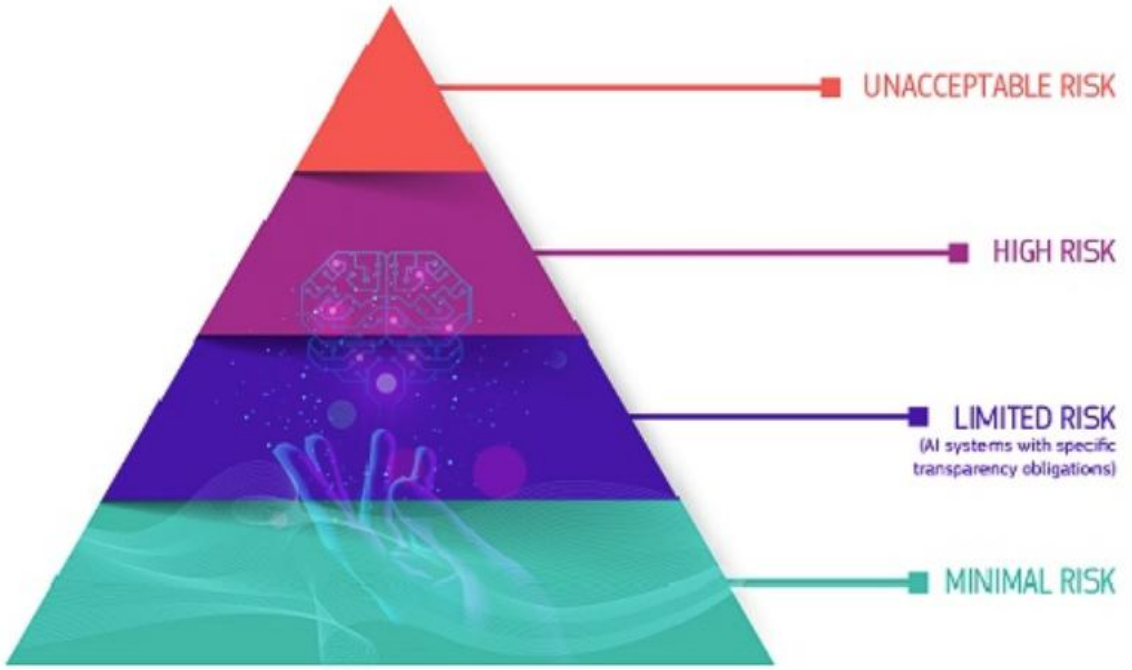
Home > Policies > Artificial Intelligence > European approach to artificial intelligence > AI Act

AI Act

The AI Act is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

A Risk-based Approach

The AI Act defines 4 levels of risk for AI systems:



Μη αποδεκτός κίνδυνος (Απαγορεύεται η σχεδίαση, ανάπτυξη και χρήση τέτοιων εργαλείων)

- Επιβλαβής χειριστικότητα και εξαπάτηση
- Επιβλαβής εκμετάλλευση ευάλωτων στοιχείων
- Κοινωνική διαστρωμάτωση/αξιολόγηση
- Πρόβλεψη/Αξιολόγηση κινδύνου ενός εγκλήματος ιδιώτη
- Μη στοχευμένη αναζήτηση στο διαδίκτυο για την δημιουργία/επέκταση βάσεων δεδομένων για αναγνώριση προσώπων
- αναγνώριση συναισθήματος σε επαγγελματικά και εκπαιδευτικά περιβάλλοντα
- Ανάλυση βιοδεικτών για την συναγωγή ιδιωτικών χαρακτηριστικών
- Αναγνώριση σε πραγματικό χρόνο βιοδεικτών εξ αποστάσεως σε δημόσιους χώρους με στόχο την επιβολή του νόμου

Υψηλός κίνδυνος (Τέτοια εργαλεία υπόκεινται σε αυστηρούς κανόνες/προϋποθέσεις)

- Συστήματα ασφαλείας σε κρίσιμες υποδομές (πχ μεταφορές), των οποίων η αποτυχία μπορεί να θέσει σε ρίσκο την υγεία και τις ζωές πολιτών
- Συστήματα ΤΝ στην εκπαίδευση, που μπορεί να καθορίζουν την πρόσβαση στην εκπαίδευση και την σταδιοδρομία
- Βασιζόμενα σε ΤΝ στοιχεία ασφάλειας σε εργαλεία (πχ ΤΝ στην εφαρμογή ρομποτικής χειρουργικής)
- Εργαλεία ΤΝ για την εύρεση εργασίας, την διαχείριση υπαλλήλων (πχ λογισμικό που αξιολογεί βιογραφικά υποψηφίων)
- Εργαλεία ΤΝ που δίνουν ή όχι πρόσβαση σε συγκεκριμένες υπηρεσίες (πχ λογισμικό που επιτρέπει η όχι την δανειοδότηση)
- Συστήματα απομακρυσμένης αναγνώρισης βιοδεικτών
- Συστήματα Τν στην επιβολή του νόμου, που μπορεί να παραβιάζουν βασικά ανθρώπινα δικαιώματα (πχ αξιολόγηση της αξιοπιστίας στοιχείων)
- Εργαλεία ΤΝ για την διαχείριση μετανάστευσης, ασύλου, συνόρων (πχ αξιολόγηση ειτήσεων για visa)
- Εργαλεία ΤΝ που χρησιμοποιούνται για δικαστικές αποφάσεις και σε διαδικασίες δημοκρατίας (πχ ετοιμάζουν αποφάσεις δικαστικές)

Περιορισμένος κίνδυνος (Κίνδυνος που αφορά Διαφάνεια)

- Το AI Act επιβάλλει την ενημέρωση των ανθρώπων όπου χρειάζεται, ώστε να διατηρείται η εμπιστοσύνη.
- Πχ όποτε χρησιμοποιείται ένας personal assistant πρέπει ο άνθρωπος να είναι ενήμερος ότι επικοινωνεί με μηχανή.
- Η παραγωγική ΤΝ πρέπει να παράγει περιεχόμενο το οποίο να ξεχωρίζει από το ανθρώπινο. Πχ τα deep fakes πρέπει να αναγράφονται ότι αποτελούν προϊόν ΤΝ.

Καθόλου ή ελάχιστος κίνδυνος

- Σε αυτή την κατηγορία ανήκει η πλειοψηφία των εφαρμογών ΤΝ.
- Δεν θέτει κανόνες για αυτή την κατηγορία το AI Act.
- Πχ video games με ΤΝ ή φίλτρα για spam.

Ερωτήματα που παίρνουμε μαζί…

- Ποιος ευθύνεται όταν εργαλεία ΤΝ βλάπτουν;
- Υπάρχουν διαφορές ανάμεσα στην Ηθική που πρέπει να επιβάλλεται στην ΤΝ και στον άνθρωπο;
- Σε τι βαθμό μπορούμε να «κλείσουμε τα μάτια» στους κινδύνους που ελλοχεύουν σε συστήματα ΤΝ, και να γευόμαστε τα πλεονεκτήματά τους;
- Τι μπορεί να κάνει ο καθένας από εμάς ως Πολίτης για να επηρεάσει ως προς τον τρόπο που χρησιμοποιείται η ΤΝ;