

Code4Thought

Δημιουργώντας Εμπιστοσύνη στην
Τεχνητή Νοημοσύνη με τη χρήση Τεχνητής Νοημοσύνης:
Βέλτιστες Πρακτικές και Παραδείγματα

21/10/2020

Γιάννης Κανελλόπουλος, Ανδρέας Μεσσαλάς

Τι μας κρατάει ξύπνιους το βράδυ

- Τις τελευταίες 2 δεκαετίες η ομάδα μας αναλύει και αξιολογεί συστήματα λογισμικού μεγάλης κλίμακας προκειμένου να βοηθήσουμε τις εταιρείες να αντιμετωπίσουν πιθανά ρίσκα και ελαττώματα που σχετίζονται με αυτά.
- Έχουμε συνειδητοποιήσει ότι η τεχνολογία που παράγει ένας οργανισμός είναι ο καθρέφτης του.
- Στην **Code4Thought** μετατρέπουμε όλη αυτή την τεχνογνωσία στην τεχνολογία PyThia που διασφαλίζει ότι τα μοντέλα AI/ML θα χαρακτηρίζονται από:
 - **Αντικειμενικότητα (Fairness)**
 - **Λογοδοσία (Accountability)**
 - **Διαφάνεια (Transparency)**

Agenda

- Περιγραφή Προβλήματος και Κανονιστικού Πλαισίου
- Ανάλυση χαρακτηριστικών F.Acc.T (Fairness, Accountability, Transparency) και προκλήσεις στα πλαίσια της οδηγίας GDPR
- Η εμπιστοσύνη ως απαραίτητο συστατικό επιτυχίας των αλγοριθμικών συστημάτων
- PyThia Demo

Πρόσφατες ειδήσεις

MIT Technology Review

Artificial Intelligence Oct 25

A biased medical algorithm favored white people for health-care programs

PUBLIC RELEASE: 9-APR-2015

Who's a CEO? Google image results can shift gender biases

Harvard
Business
Review

TECHNOLOGY

UNIVERSITY OF WASHINGTON



When Algorithms Decide Whose Voices Will Be Heard

by Theodora (Theo) Lau and Uday Akkaraju



Artificial Intelligence Nov 11

Apple Card is being investigated over claims it gives women lower credit limits

Artificial Intelligence Dec 20

A US government study confirms most face recognition systems are racist

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

THE VERGE

TECH

REVIEWS

SCIENCE

CREATORS

ENTERTAINMENT

VIDEO

MORE

f t

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent |

Artificial Intelligence

There's an easy way to make lending fairer for women. Trouble is, it's illegal.

Goldman Sachs defended itself in the Apple Card scandal by saying it did not consider gender when calculating creditworthiness. If it did, that could actually mitigate the problem.

by Karen Hao | a month ago

Η τεχνολογία σαν μέρος της ιστορίας

Microsoft Urged To Follow Amazon And IBM: Stop Selling Facial Recognition To Cops After George Floyd's Death



Thomas Brewster Forbes Staff

Cybersecurity

Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

f

t

in



Technology

Microsoft won't sell police its facial-recognition technology, following similar moves by Amazon and IBM

Big tech companies back away from selling facial recognition to police. That's progress.

After IBM, Amazon, and Microsoft upend their facial recognition businesses, attention turns to federal lawmakers.

By **Rebecca Heilweil** | Updated Jun 11, 2020, 5:02pm EDT

Κανονιστικά πλαίσια και Νομοθεσία

- Γενικός Κανονισμός για την Προστασία Δεδομένων ([GDPR](#)), Ευρωπαϊκή Ένωση (Απρίλιος, 2016)
- Personal Information Protection and Electronic Documents Act ([PIPEDA](#)), Καναδάς (2001)
- California Consumer Privacy Act ([CCPA](#)), Η.Π.Α (Ιούνιος, 2018)
- [White paper](#) on Artificial Intelligence - A European approach to excellence and trust (Φεβρουάριος, 2020)
- [Νομοθεσία](#) στη Δανία για AI & Data Ethics (Μάιος, 2020)

Αρχές της GDPR

Άρθρο 5 GDPR:

- (1) “Τα δεδομένα προσωπικού χαρακτήρα υποβάλλονται σε σύννομη και θεμιτή επεξεργασία με διαφανή τρόπο σε σχέση με το υποκείμενο των δεδομένων”
 - νομιμότητα (lawfulness)
 - **αντικειμενικότητα (fairness)**
 - **διαφάνεια (transparency)**
 - ...

- (2) “Ο υπεύθυνος επεξεργασίας φέρει την ευθύνη και είναι σε θέση να αποδείξει τη συμμόρφωση με την παράγραφο 1”
 - **λογοδοσία (accountability)**

Αντικειμενικότητα και GDPR (Fairness)

Πληθώρα Ερμηνειών

- Μετάφραση του fair στα Ελληνικά στη GDPR:
 - Στα άρθρα: **θεμιτή**
 - Στις Αιτιολογικές Σκέψεις (recitals): **δίκαιη**
- Διάφορες ερμηνείες:
 - Μη διάκριση
 - **Δίκαιη εξισορρόπηση** [1],
 - αναλογικότητα μεταξύ συμφερόντων και αναγκαιότητας σκοπών,
 - **Διαδικαστική δικαιοσύνη** [1],
 - καθήκοντα διαφάνειας, την επικαιρότητα και το βάρος της φροντίδας των υπευθύνων επεξεργασίας δεδομένων.

Σύνδεση διαφάνειας-μη διάκρισης-αντικειμενικότητας

Άρθρο 13(2), 14(2) και 40(2):

“για την εξασφάλιση θεμιτής και διαφανούς επεξεργασίας”

Αιτιολογική Σκέψη 71:

“Προκειμένου να διασφαλισθεί δίκαιη και διαφανής επεξεργασία σε σχέση με το υποκείμενο των δεδομένων, [...], ο υπεύθυνος επεξεργασίας θα πρέπει [...] να εφαρμόζει τεχνικά και οργανωτικά μέτρα, ώστε να διορθώνονται οι παράγοντες που οδηγούν σε ανακρίβειες σε δεδομένα προσωπικού χαρακτήρα και να ελαχιστοποιείται ο κίνδυνος σφαλμάτων,” όπως “τα αποτελέσματα διακρίσεων σε βάρος φυσικών προσώπων βάσει της φυλετικής ή εθνοτικής καταγωγής, των πολιτικών φρονημάτων, της θρησκείας ή των πεποιθήσεων, [...]”

Προστατευόμενες Ομάδες

- **Φυλή και Εθνική καταγωγή, Χρώμα**
 - *[Directive 2000/43/EC, Council's Framework Decision on combating racism and xenophobia],*
 - *[Civil Rights Act of 1964]*
- **Θρησκεία ή πίστη, αναπηρία, ηλικία, σεξουαλικός προσανατολισμός**
 - *[Directive 2000/78/EC, Directive Proposal (COM(2008)462), Charter of Fund. Rights (Art. 21)]*
 - *[Civil Rights Act of 1964, Equal Pay Act of 1963, Age Discrimination in Employment Act of 1967]*
- **Φύλο**
 - *[Charter of Fundamental Rights (Art. 21 & 23), Directive 2006/54/EC, TFEU (Art. 8 157)]*
 - *[Civil Rights Act of 1964, Equal Pay Act of 1963]*
- **Οικογενειακή κατάσταση**
 - *[Charter of Fundamental Rights, Art. 7]*
 - *[Civil Rights Act of 1968]*

● ...

Πολλές Μετρικές Αντικειμενικότητας

Translation tutorial: 21 fairness definitions and their politics

Arvind Narayanan
@random_walker

0:07 / 55:20

Tutorial: 21 fairness definitions and their politics

Άποψη της Code4Thought: Μετρική Αντικειμενικότητας

Disparate Impact Ratio

- Το ποσοστό των ατόμων που λαμβάνουν ένα συγκεκριμένο αποτέλεσμα για δύο ομάδες.
- Κανόνας των $\frac{4}{5}$ [U.S. Equal Employment Opportunity Commission (EEOC)]
 - *“a selection rate for any race, sex, or ethnic group less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded [...] as evidence of adverse impact” [3]*
- Industry standard

Διαφάνεια και GDPR (Transparency)

Διαμάχη στα πλαίσια της GDPR

Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation FREE

Sandra Wachter, Brent Mittelstadt, Luciano Floridi [Author Notes](#)

International Data Privacy Law, Volume 7, Issue 2, May 2017, Pages 76–99,

<https://doi.org/10.1093/idpl/ix005>

Published: 03 June 2017

Πηγή: [4]

Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation

Gianclaudio Malgieri [✉](#), Giovanni Comandé

International Data Privacy Law, Volume 7, Issue 4, November 2017, Pages 243–265,

<https://doi.org/10.1093/idpl/ix019>

Published: 13 November 2017 [Article history ▼](#)

Πηγή: [5]

Δεσμευτική απόφαση

Άρθρο 22 παράγραφος 3 GDPR

“ο υπεύθυνος επεξεργασίας εφαρμόζει κατάλληλα μέτρα για την προστασία των δικαιωμάτων, των ελευθεριών και των έννομων συμφερόντων του υποκειμένου των δεδομένων, τουλάχιστον του δικαιώματος εξασφάλισης ανθρώπινης παρέμβασης από την πλευρά του υπευθύνου επεξεργασίας, **έκφρασης άποψης και αμφισβήτησης της απόφασης.**”.

Μη δεσμευτική απόφαση

Αιτιολογική σκέψη (recital) 71 GDPR

“Το υποκείμενο των δεδομένων θα πρέπει να έχει το δικαίωμα να μην υπόκειται σε απόφαση, [...] λαμβανόμενη αποκλειστικά βάσει αυτοματοποιημένης επεξεργασίας [...] όπως

- η αυτόματη άρνηση αίτησης πίστωσης ή
- πρακτικές ηλεκτρονικών προσλήψεων χωρίς ανθρώπινη παρέμβαση.

Ωστόσο, θα πρέπει να επιτρέπεται [...] όταν το υποκείμενο των δεδομένων παρέχει τη ρητή συγκατάθεσή του. Σε κάθε περίπτωση, η επεξεργασία αυτή θα πρέπει να υπόκειται σε κατάλληλες εγγυήσεις, [...] όπως το δικαίωμα να λάβει αιτιολόγηση της απόφασης που ελήφθη στο πλαίσιο της εν λόγω εκτίμησης”

Η διαμάχη δεν τελειώνει εδώ...

Άρθρο 15(1) (η) GDPR:

ο υπεύθυνος επεξεργασίας πρέπει να παρέχει

“σημαντικές πληροφορίες σχετικά με τη λογική που ακολουθείται”

“πληροφορίες σχετικά με τη λογική που ακολουθείται”

- Μερικοί ακαδημαϊκοί ([4], [5], [6]) επιχειρηματολογούν ότι:
 - αναφέρεται μόνο στη γενική δομή του μοντέλου επεξεργασίας, άρα
 - οι εξηγήσεις για ατομικές αποφάσεις δεν είναι απαραίτητες
- Άλλοι [7] πιστεύουν ότι :
 - Οι πληροφορίες μπορούν να έχουν νόημα μόνο εάν βοηθούν τα υποκείμενα των δεδομένων να ασκήσουν τα δικαιώματά τους βάσει του Άρθρο 22 (1) και (3) GDPR, το οποίο περιλαμβάνει το δικαίωμα αμφισβήτησης της απόφασης.
 - Οι εξηγήσεις των αποφάσεων παρέχουν την δυνατότητα αμφισβήτησης της απόφασης.

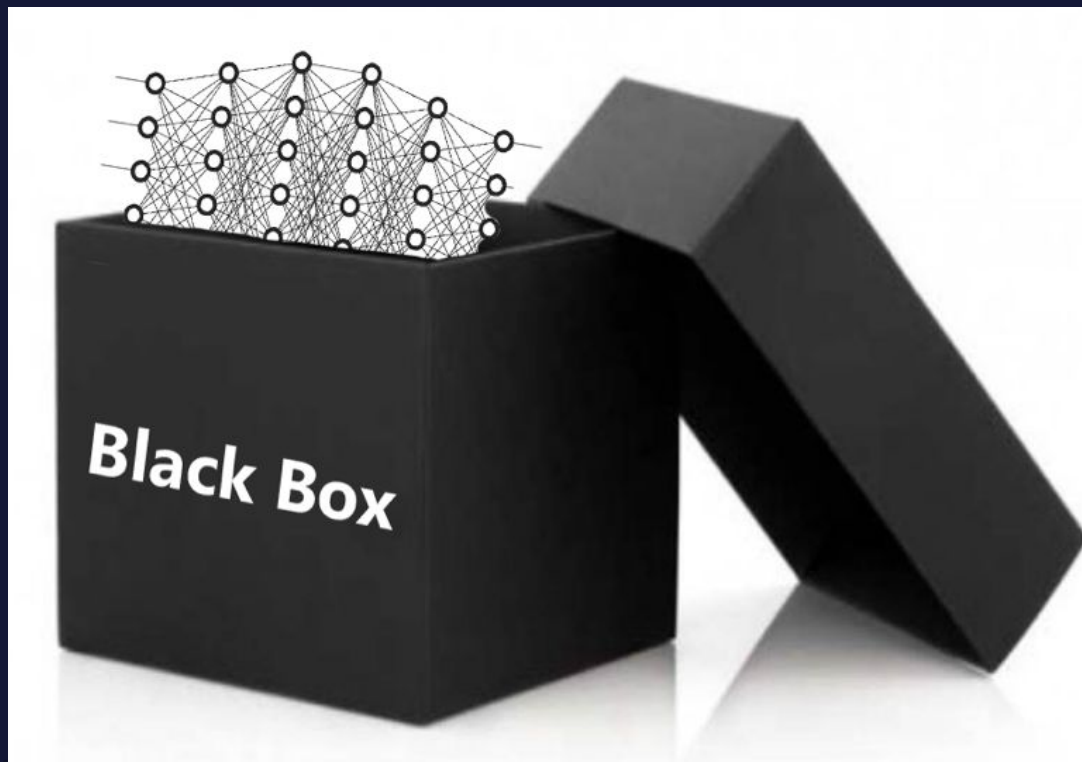
Ευκολη συμμόρφωση (“παραθυράκι”)

- “αποφάσεις αποκλειστικά βάσει αυτοματοποιημένης επεξεργασίας”
- Επομένως, κάθε είδους ανθρώπινη παρέμβαση στην εξαγόμενη από το αλγοριθμικό σύστημα απόφαση εξαιρεί τον ελεγκτή από τους περιορισμούς του Άρθρου 22 και 15 (1)(η) της GDPR

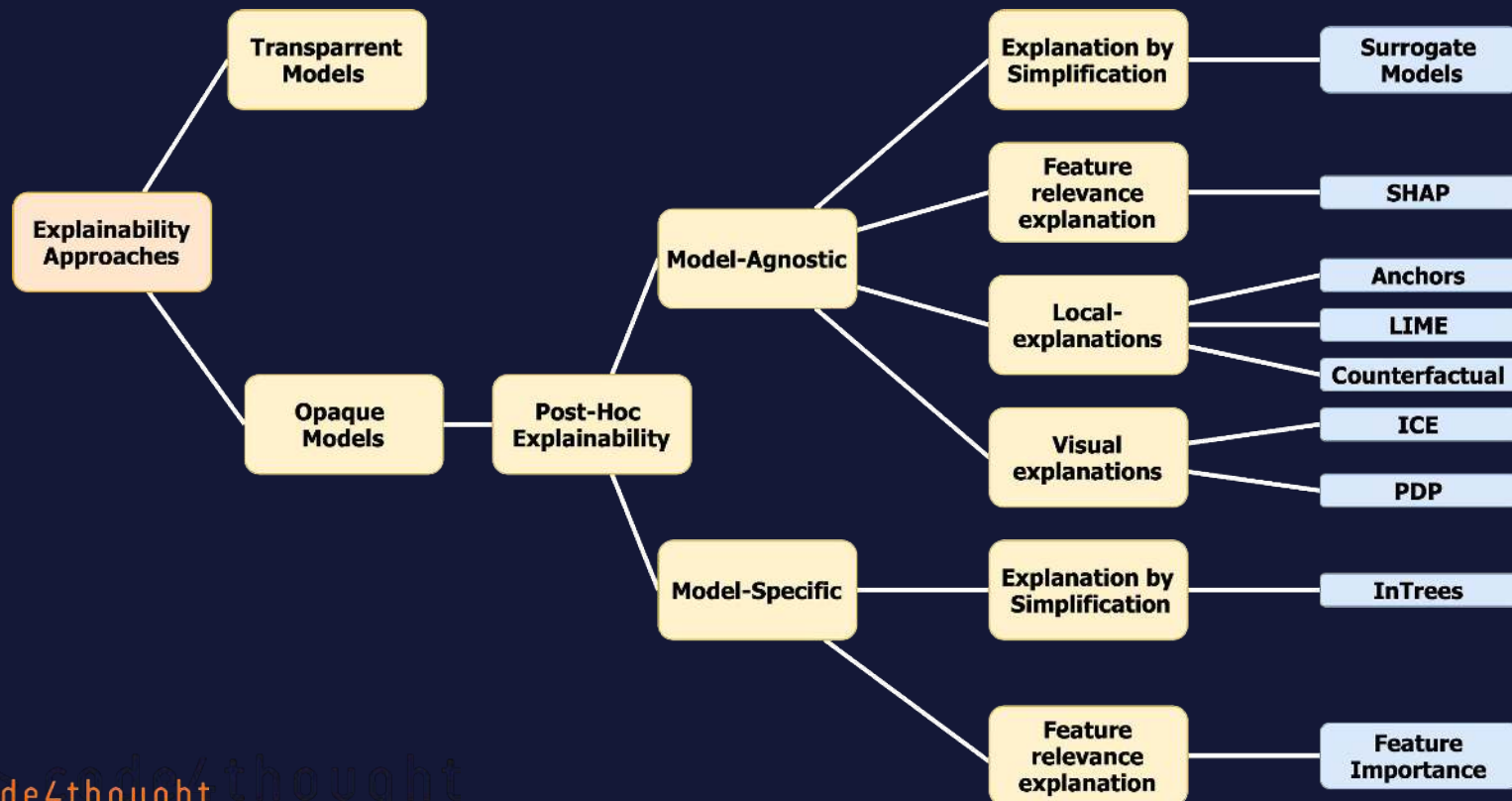
Άποψη της Code4Thought

- Η παροχή της εξήγησης πρέπει να απαιτείται σε κάθε περίπτωση αυτόματης αλγοριθμικής απόφασης ακόμη και αν διαμεσολαβεί ανθρώπινος παράγοντας για τη λήψη της τελικής απόφασης.
- Η παροχή της εξήγησης πρέπει πρωτίστως να γίνεται σε ατομικό επίπεδο (local explanation) και στη συνέχεια σε γενικό/ολικό (global explanation).
- Η χρήση της τεχνολογίας μπορεί να διευκολύνει και αυτοματοποιήσει την παροχή εξηγήσεων σε κάθε επίπεδο (ατομικό/γενικό).

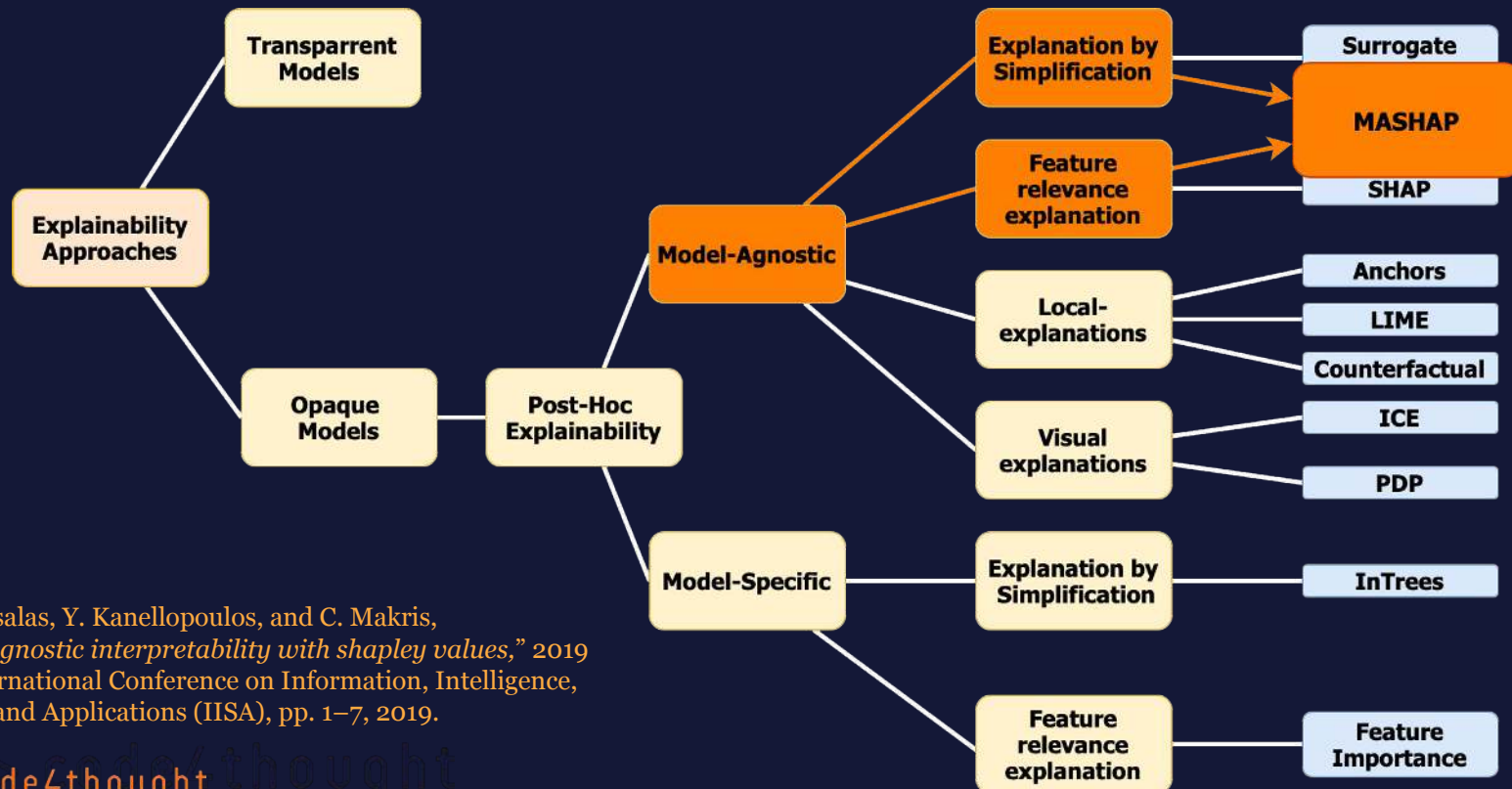
Τεχνητή Νοημοσύνη ως “Μαύρο Κουτί”



Κατηγορίες μεθόδων επεξήγησης



Κατηγορίες μεθόδων επεξήγησης



* A. Messalas, Y. Kanellopoulos, and C. Makris,
“Model-agnostic interpretability with shapley values,” 2019
10th International Conference on Information, Intelligence,
Systems and Applications (IISA), pp. 1–7, 2019.

Λογοδοσία (Accountability)

Η εξουσία εκφράζεται όλο και περισσότερο αλγοριθμικά

“Ηδη σήμερα, η “αλήθεια” καθορίζεται
από τα κορυφαία αποτελέσματα
της αναζήτησης στο Google.”

Yuval Noah Harari,
“21 μαθήματα για τον 21ο αιώνα”

Λογοδοσία και GDPR

Άρθρο 5 (2) GDPR:

“Ο υπεύθυνος επεξεργασίας φέρει την ευθύνη και είναι σε θέση να αποδείξει τη συμμόρφωση με την παράγραφο 1 (λογοδοσία)”

Άρθρο 24 (1) GDPR:

“ο υπεύθυνος επεξεργασίας εφαρμόζει κατάλληλα τεχνικά και οργανωτικά μέτρα προκειμένου να διασφαλίζει και να μπορεί να αποδεικνύει ότι η επεξεργασία διενεργείται σύμφωνα με τον παρόντα κανονισμό.”

Αξιολογώντας τη Λογοδοσία*: Χρήση Ερωτηματολογίου (checklist)

The screenshot shows the Code4Thought dashboard with a sidebar containing 'DASHBOARD' and 'MY MODELS'. The main area displays four audit checklist items:

Item	Status	Dropdown
No Properties here	Red X	> Algorithm
Algorithmic Presence	Yellow Warning Triangle	> Algorithm
Accuracy	Red X	> Organization
Auditability Fairness	Yellow Warning Triangle	> Organization

Callouts from the image:

- Unsupervised model**: Points to the 'No Properties here' item.
- Not priorities**: Points to the 'Accuracy' and 'Auditability Fairness' items.
- No annotations**: Points to the 'Accuracy' item.

* Y. Kanellopoulos, "Accountability of Algorithmic Systems: How We Can Control What We Can't Exactly Measure" [Cutter Business Technology Journal](#), March 2019.

** H. Tagiou, Y. Kanellopoulos, C. Makris, C. Aridas, "A tool supported framework for the Assessment of Algorithmic Accountability", in *International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2019.

*** High-Level Expert Group on Artificial Intelligence (AI HLEG), "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment", 2020

Χτίζοντας Εμπιστοσύνη

Προκλήσεις για επιτυχή υλοποίηση AI/ML

- Επιλογή σωστής λύσης (δηλ. κατάλληλο μοντέλο) για ένα δεδομένο επιχειρηματικό πρόβλημα.
- Δημιουργία κατάλληλων συνόλων δεδομένων εκπαίδευσης (π.χ. έλλειψη ετικετών) για τα μοντέλα.
- Έλλειψη εμπιστοσύνης στα αποτελέσματα ενός μοντέλου κατά την ανάπτυξη του.

Προκλήσεις για χτίσιμο εμπιστοσύνης

- Οι τεχνικές ομάδες πασχίζουν για ακρίβεια και γρήγορη παράδοση και όχι τόσο για τη δημιουργία εμπιστοσύνης.
- Η λογοδοσία και η αντικειμενικότητα είναι απλώς μεταγενέστερες σκέψεις από τη λειτουργική παράδοση του μοντέλου.
- Όταν επιβάλλεται εμπιστοσύνη ως κανονιστική απαίτηση (π.χ. διαφάνεια) εφαρμόζονται ad-hoc και εφάπαξ λύσεις. Δεν υπάρχει συνέχεια, βιωσιμότητα των λύσεων.

Αρχές της Code4Thought για την PyThia

- Απλότητα αλλά όχι απλοϊκότητα.
- Διαφάνεια αλλά με επιλεκτικότητα.
- Χρήση αναφορών/προτύπων/ερωτηματολογίων.

Αναφορές

- [1] D. Clifford and J. Ausloos, "Data Protection and the Role of Fairness", *Yearbook of European Law*, vol. 37, pp. 130-187, 2018.
- [2] Narayanan A, Tutorial: 21 fairness definitions and their politics, ACM FAT*, 2018 [[video](#)]
- [3] EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (2018).
- [4] Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int Data Priv Law* 7(2):76-99
- [5] Malgieri G, Comandé G (2017) Why a right to legibility of automated decision-making exists in the general data protection regulation. *Int Data Priv Law* 7(6):243-265
- [6] Party ADPW (2017) Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679
- [7] Selbst AD, Powles J (2017) Meaningful information and the right to explanation. *Int Data Priv Law* 7(4):233-242

Το επιστημονικό υπόβαθρο της PyThia

- Βιβλία:

- *“Use model-agnostic explanations for finding bias in black box models”, Y. Kanellopoulos and A. Messalas, “97 Things About Ethics” edited by Bill Franks, ISBN-13: 978-1492072669, O’Reilly, August 2020;*
- *“Make Accountability a Priority”, Y. Kanellopoulos, “97 Things About Ethics” edited by Bill Franks, ISBN-13: 978-1492072669, O’Reilly, August 2020.*

- Ακαδημαϊκά Συνέδρια:

- *“PyThia: A Reporting Tool on Bias Evaluation and Mitigation”, Mechanism Design For Social Good International Workshop (MD45G), August 2020;*
- *“Model-Agnostic Interpretability with Shapley Values” A. Messalas, Y. Kanellopoulos, C. Makris. International Conference on Information, Intelligence, Systems and Applications (IISA), July 2019;*
- *“A tool supported framework for the Assessment of Algorithmic Accountability”, H. Tagiou, Y. Kanellopoulos, C. Makris, C. Aridas, International Conference on Information, Intelligence, Systems and Applications (IISA), July 2019.*

- Περιοδικά (Journals):

- *“Accountability of Algorithmic Systems: How We Can Control What We Can’t Exactly Measure”, Yiannis Kanellopoulos, Cutter Business Technology Journal, March 2019.*

PyThia Demo