Ориентиране в етика на ИИ

Образователен форум „Образованието през парадигмата на хуманизма и изкуствения интелект“, Варна

ДИМИТЪР НИКОЛОВ, Д-Р
d_nikolov@tu-sofia.bg

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ СОФИЯ
ФЕТТ

АДРЕС
бул. Климент Охридски 8, гр. София

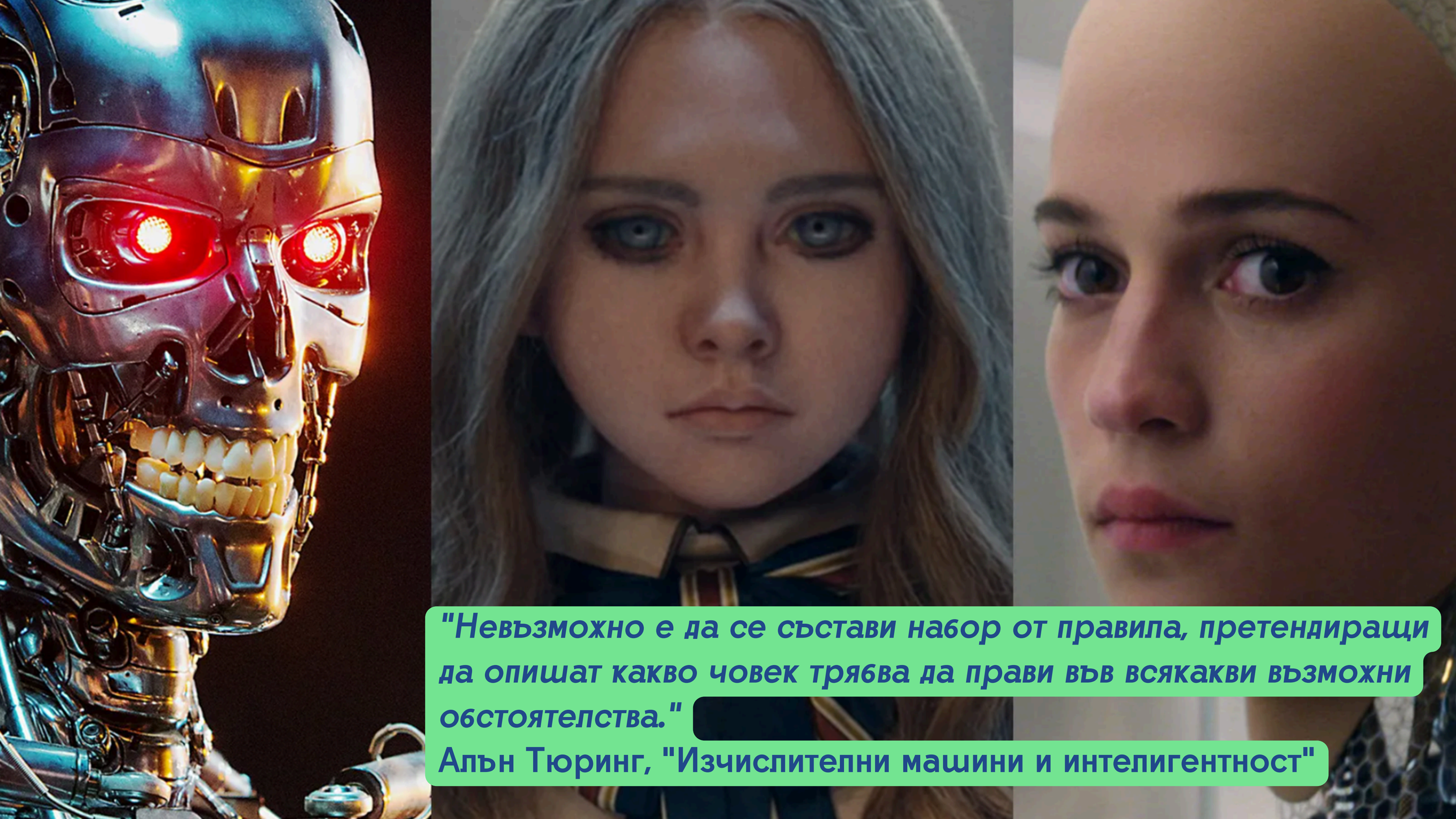
За мен

доц. д-р Димитър Николов

- ФЕТТ при ТУ-София
- Микро Нано Лаборатория, СТП АД - ръководител
- Клъстер Микроелектроника и индустриалнни електронни системи - председател на УС
- AI клъстер България
- 50+ публикации
- 15 Европейски проекти

d_nikolov@tu-sofia.bg

Съдържание

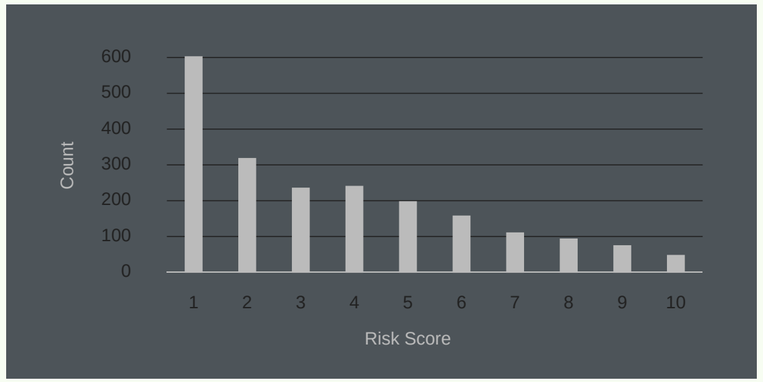


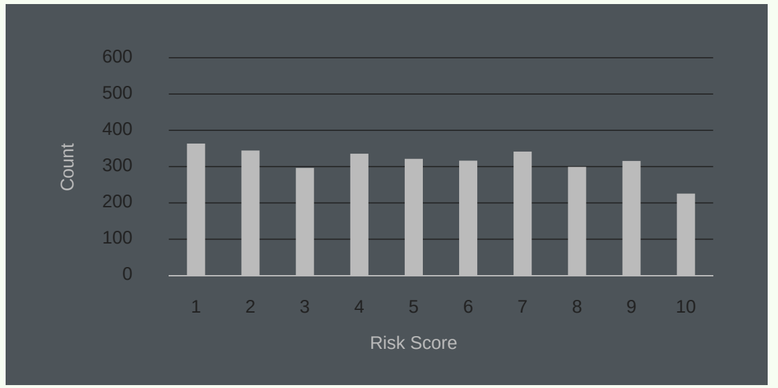
"Невъзможно е да се състави набор от правила, претендиращи да опишат какво човек трябва да прави във всякакви възможни обстоятелства."

Алън Тюринг, "Изчислителни машини и интелигентност"

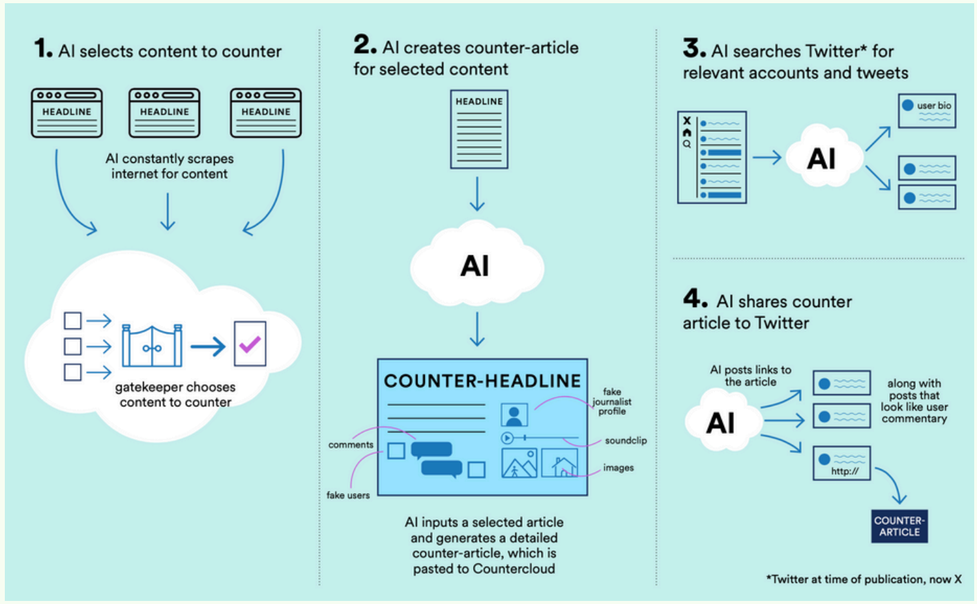
Propublica <-> COMPASS

Концепцията за справедливост в изкуствения интелект включва равноправно отношение към индивидите от различни демографски групи.
Справедливостта е от съществено значение в области с висок риск като правосъдие, медицина, изборен процес, заетост, енергетика, финанси.
Въпреки усилията, много системи базирани на изкуствен интелект показват предразсъдъци и водят до дискриминация.





Сравнение между оценката на риска



Употреба в търсене на влияние в изборния процес

J. Angwin, S. Mattu, J. Larson, & L. Kirchner. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Nestor Maslej, et. al, "The AI Index 2024 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024

Етични норми и тяхното влияние 1

Какво е реалното въздействие на етичните насоки върху човешкото вземане на решения в областта на изкуствения интелект и машинното обучение?

European Commission's Ethics Guidelines for Trustworthy AI (2019)
Фокусира се върху осигуряване на законност, етика и устойчивост на ИИ системите.

OECD Principles on AI (2019)
Подчертава необходимостта от растеж, устойчивото развитие и благосъстоянието.

Beijing AI Principles (2019)
Призовава за отговорна употреба на ИИ за насърчаване на социалното добро.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019)
Предоставя цялостна рамка за етични съображения в ИИ.

AI Now Report (2018)
Подчертава необходимостта от отговорност и прозрачност в ИИ системите.

Partnership on AI (2018)
Сътрудничество между основни технологични компании за насърчаване на добри практики в ИИ.

Montreal Declaration for Responsible AI (2017)
Цели да ръководи развитието на ИИ по начин, който уважава правата на човека.

Asilomar AI Principles (2017)
Фокусира се върху дългосрочната безопасност и етичните съображения на ИИ.

Future of Life Institute's AI Principles (2017)
Призовава за безопасна и полезна употреба на ИИ технологиите.

UK House of Lords AI Report (2018)
Обсъжда последствията от ИИ за обществото и необходимостта от регулация.

AI Ethics Guidelines by the Canadian Government (2019)
Подчертава прозрачността, отговорността и справедливостта в ИИ системите.

Singapore's Model AI Governance Framework (2019)
Предоставя рамка за отговорно управление на ИИ.

UNESCO's Recommendation on the Ethics of AI (2021)
Цели да насърчи глобалното етично развитие на ИИ.

World Economic Forum AI Principles (2018)
Фокусира се върху социалното въздействие на ИИ и необходимостта от етични рамки.

OpenAI Charter (2018)
Очертава принципи за отговорно развитие на ИИ технологиите.

UAI4People's Ethical Framework (2018)
Предлага рамка за етичен ИИ, която приоритизира благосъстоянието на хората.

The Montreal AI Ethics Institute's Guidelines (2019)
Цели да насърчи етичните практики в ИИ чрез ангажираност на общността.

The Data Ethics Framework by the UK Government (2018
Предоставя насоки за етично използване на данни в ИИ приложения.

The AI Ethics Guidelines by the Australian Government (2020)
Фокусира се върху отговорната употреба на ИИ в публичната политика.

The European Parliament's Resolution on AI (2020)
Призовава за цялостна регулаторна рамка за ИИ технологиите.

The AI Ethics Guidelines by the German Government (2020
Подчертава важността на човешкия надзор в ИИ системите.

The AI Ethics Guidelines by the French Government (2020)
Призовава за етичен ИИ, който уважава основните права и свободи.

Етични норми и тяхното влияние 2

Имат ли тези етични насоки реално въздействие върху човешкото вземане на решения в областта на изкуствения интелект и машинното обучение?

Общи насоки в етичните норми

1. Отговорност
2. Прозрачност
3. Справедливост и равенство
4. Човекоцентричен дизайн
5. Поверителност и защита на данните
6. Безопасност и сигурност
7. Сътрудничество и ангажираност на заинтересованите страни
8. Непрекъснато наблюдение и оценка

Настоящите етични насоки рядко оказват значително влияние върху вземането на решения в AI индустрията. Често тези насоки служат като PR инструмент, без реално да променят практиките на компаниите.

Основни аспекти

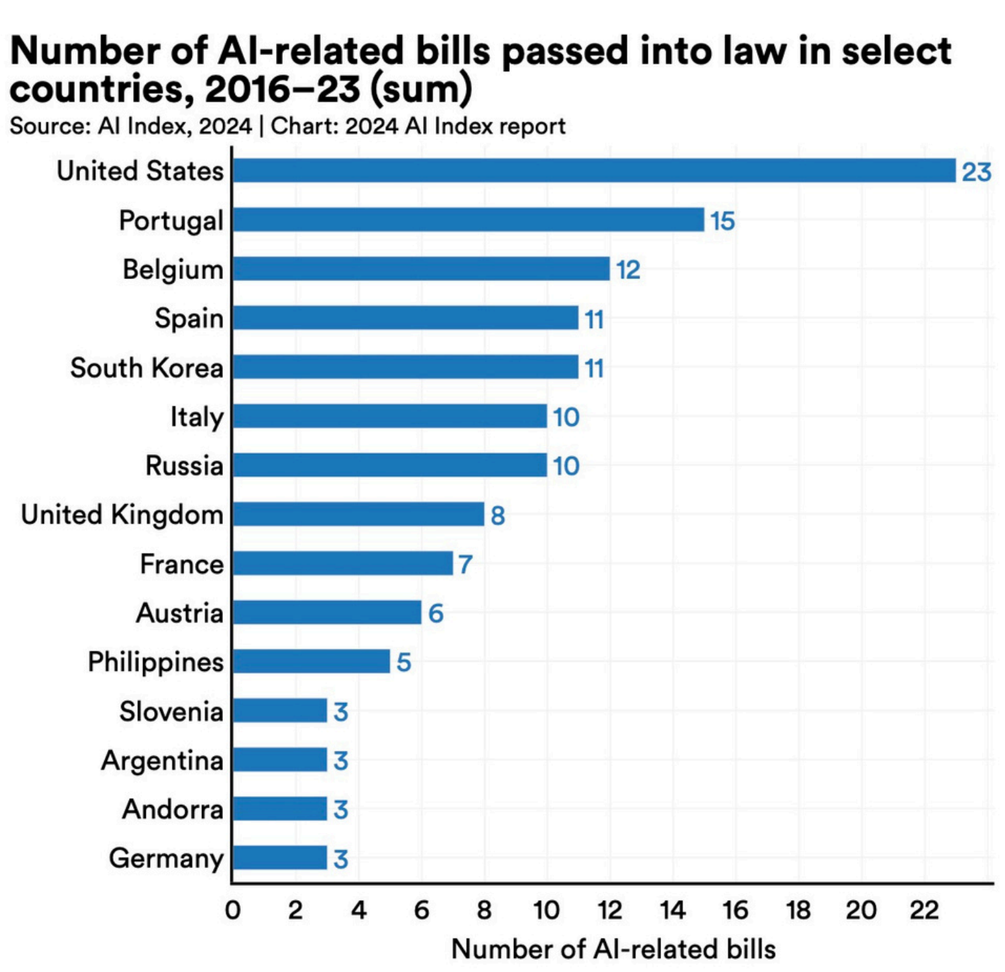
1. Ефективност на етичните насоки в AI.
2. Пропуски в съществуващите насоки.
3. Технически и етични предизвикателства.
4. Препоръки за подобряване на насоките.

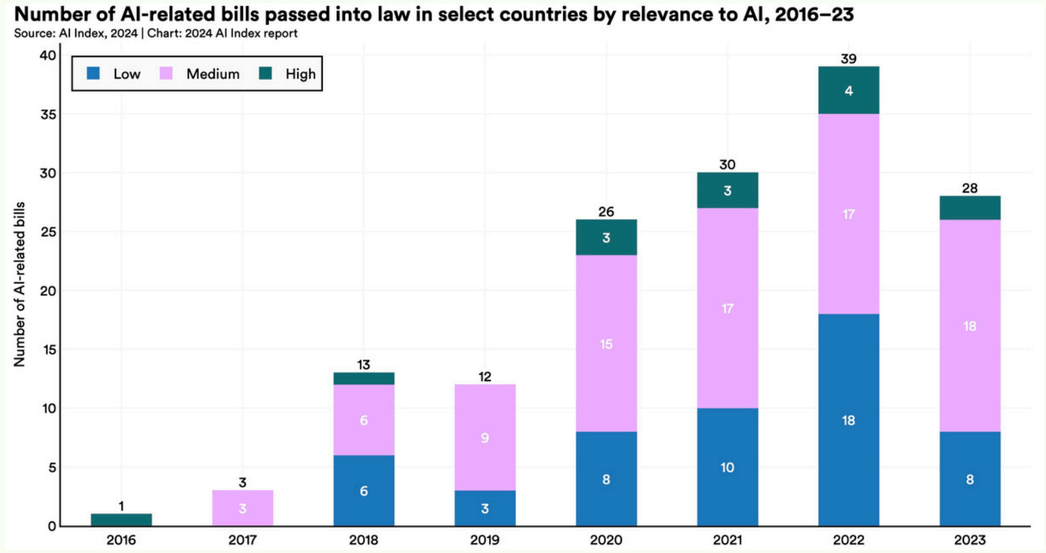
Конфликтни точки

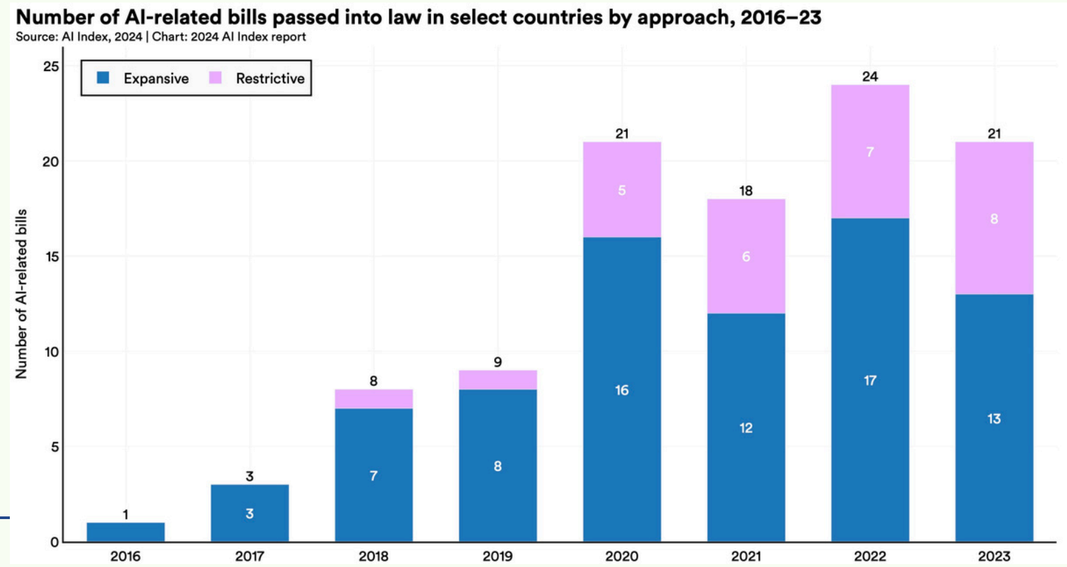
1. Точност <-> справедливост
2. Точност <-> обяснимост
3. Поверителност <-> прозрачност
4. Качество на услугите <-> поверителност
5. Персонализация <-> солидарност
6. Удобство <-> достойнство
7. Ефикасност <-> безопасност и устойчивост
8. Удовлетвореност <-> равенство

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8
Whittlestone, J., et. al. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Nuffield Foundation. https://www.nuffieldfoundation.org/wp-content/uploads/2019/02/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf
Weinberg, L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. https://doi.org/10.48550/ARXIV.2205.04460

Законодателство 2016 - 2023 географско покритие

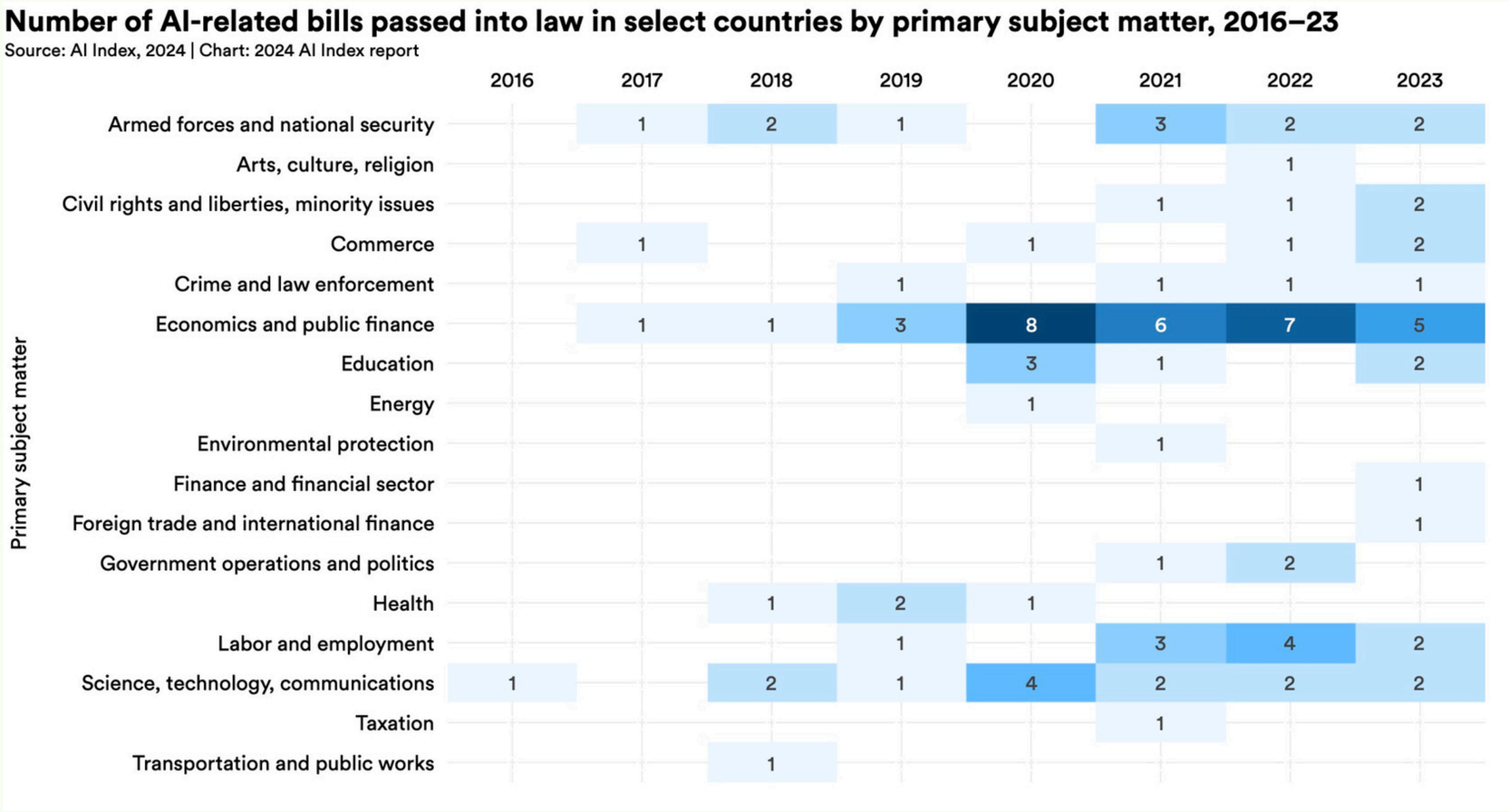






NESTOR MASLEJ, ET. AL, “THE AI INDEX 2024 ANNUAL REPORT,” AI INDEX STEERING COMMITTEE, INSTITUTE FOR HUMAN-CENTERED AI, STANFORD UNIVERSITY, STANFORD, CA, APRIL 2024

Законодателство 2016 - 2023 по тематични области



NESTOR MASLEJ, ET. AL, “THE AI INDEX 2024 ANNUAL REPORT,” AI INDEX STEERING COMMITTEE, INSTITUTE FOR HUMAN-CENTERED AI, STANFORD UNIVERSITY, STANFORD, CA, APRIL 2024

Timeline of the EU AI Act

Violations of the AI Act could draw **fines** of up to €35 million ($38 million USD), or 7% of a company's global revenue.

Apr 2019
EU Ethics Guidelines for Trustworthy AI

Jun 2023
EU AI Act Published

Dec 2023
Parliament reaches provisional agreement

Feb 2024
EU countries strike deal on AI rulebook

March 13, 2024
EU lawmakers give final approval to EU AI Act

May '24
EU AI Act officially became law

6 months
Nov/Dec '24
Bans on prohibited practices enter the lawbooks

9 months
Feb '25
Codes of conduct will be applied

12 months
May '25
Rules for general purpose AI systems like chatbots become applicable.

24 months
2026
Rules for high-risk AI systems (with some exceptions)

36 months
2027
Regulations for all systems come into force*

* high-risk models must undergo 3rd party conformity assessment

High Risk

Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Minimal Risk

All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

Unacceptable Risk

Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

Limited Risk

Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

On the 17 of July 2020, the High-Level Expert Group on Artificial Intelligence (AI HLEG) presented their final Assessment List for Trustworthy Artificial Intelligence.

Shaping Europe's digital future

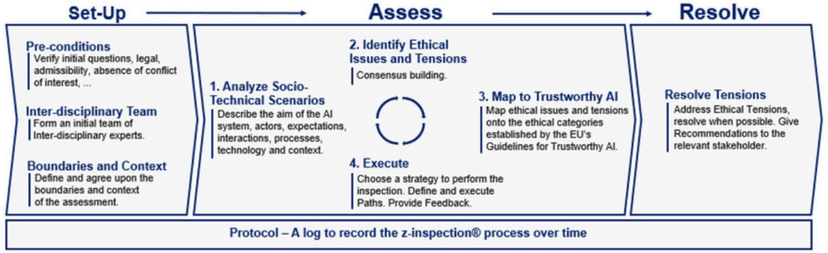


Fig. 1 Z-Inspection® process flow chart describing the main steps of the *set-up*, *assess*, and *resolve* phases. In parallel to the phases, a log is kept in which the process and events of the assessment are tracked. Adapted from Zicari et al. (2021b)

Table of Contents

https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

2024

Време за дискусия!