Artificial Intelligence System Engineering

Recognizing bias in AI systems. Estimation of bias effect on the AI system. Strategies for fair AI development and bias mitigation

Dirbtinio intelekto sistemų inžinerija

AI sistemų šališkumo atpažinimas. AI sistemos šališkumo įvertinimas. Sąžiningo AI plėtros ir šališkumo mažinimo strategijos

Bias in AI

AI systems that produce biased results that reflect and perpetuate human biases within a society, including historical and current social inequality

Bias in data is an inconsistency with the phenomenon that data represents. This inconsistency may occur for a number of reasons

Human bias

- Algorithmic Bias
- Human Bias
- Automation Bias
- Interaction Bias

Human bias

- Action-Oriented Biases
  - Action Bias
  - Illusion of control
  - Optimism Bias
  - Overconfidense Bias
- Stability Biases
- Pattern-Recognition Biases
- Interest Biases
- Social Biases

Data bias (1)

- Data Bias
  - Selection bias is the tendency to skew your choice of data sources to those that are easily available, convenient, and/or cost-effective
  - Self-selection bias - occurs when data is collected from sources that have voluntarily chosen to participate or provide information

Data bias (2)

- Data Bias
  - Omitted variable bias - featurized data doesn’t have a feature necessary for accurate prediction
  - Sponsorship or funding bias affects the data produced by a sponsored agency
  - Sampling bias (also known as distribution shift)
  - Prejudice or stereotype bias
  - Systematic value distortion is bias usually occurring with the device making measurements or observations
  - Experimenter bias
  - Labeling bias happens

Sources of algorithmic bias

- Three major sources of algorithmic bias introduced by the judgment of data scientists:
  - Confirmation bias - sets up the model to replicate a bias in the data scientist's own mind
  - Ego depletion that distracts the data scientist from opportunities to avoid bias;
  - Overconfidence that causes the data scientist to reject signals that the model might be biased.

Algorithmic Bias (1)

- Algorithmic Bias
  - as data changes, models might need some adjustments
  - A system generates outputs that are unfair or systematically prejudiced toward certain individuals or groups

Algorithmic Bias (2)

- Causes of Stability Bias:
  - System Instability: algorithms rely on relationships within a sample that may no longer hold in the future
  - Slow Response Speed:
    - Algorithms fail to adapt quickly to sudden changes in data or external factors.
    - Leads to decisions based on outdated or irrelevant patterns

Algorithmic Bias (3)

How to solve it?

- Enhancing Responsiveness:
  - Use algorithms capable of real-time learning and adjustments.
- Dynamic Data Integration:
  - Incorporate up-to-date data to minimize reliance on outdated patterns.
- Regular Validation:
  - Continuously validate models to check if assumptions align with current realities.

Biases by the algorithm

- Algorithmic Error
- Small Sample Sizes
- Tree-Based Models

Detecting algorithmic biases

- Monitor algorithms
  - Monitoring ensures early detection of biases
  - Differentiates between real-world and algorithm-introduced biases
  - Challenges with machine learning models and updates
- Metrics
  - Forward-Looking Metrics:
    - Distribution Analysis
    - Override Analysis
  - Backward-Looking Metrics:
    - Calibration Analysis
    - Rank Ordering

Detecting algorithmic biases

- Root Cause Analysis
  - Distribution analysis and decision trees
  - Hypothesis-driven approach for efficient investigation
- Black Box Challenges
  - Complexity and transparency issues
  - Continuous updates require adaptive monitoring frameworks
- Monitoring Self-Improving Algorithms
  - Rigorous tracking of model changes
  - Automated mechanisms for identifying deviations

How to tackle biases?

- Model Design
  - Get real-life insights from direct sources
  - Consider whether the data available to you now might be tainted by a real-life bias, and how best to address it
  - Ensure that the overall business design prevents fatal feedback loops
- Data Engineering
  - Be very careful about sample definition
  - Data collection - extreme caution with queries or manual processes that could unintentionally exclude or overlook certain profiles—whether due to technical glitches or errors.
  - Splitting the sample properly
  - Data quality
  - In data aggregation, test the conceptual soundness by carefully examining a couple of test cases, especially for exceptions

How to tacle biases?

- Model Assembly
  - Exclusion of records
  - Feature development
  - Short-listing
  - Model estimation
  - Model tuning
  - Calibration
  - Model documentation

How to tacle biases?

- Model Validation
  - Designating someone to challenge decisions or algorithms ensures that biases are identified and addressed
  - Validation teams often uncover significant issues in algorithms, such as hidden biases or technical flaws
  - In some organizations, especially in finance where validation is legally required, overly strict validation can become counterproductive
  - Overly strict validation processes can discourage data scientists, causing them to avoid innovation or take unnecessary precautions to pass validation
- Model Implementation
  - Implementation of the predictive features
  - Ongoing generation of fresh, unbiased data

Estimate Bias impact

- **Metrics to Evaluate Bias**:
  - False positive/negative rates
  - Disparate impact ratio
  - Equalized odds
- **Assessing the Effect**:
  - Quantifying bias in model predictions
  - Evaluating fairness across demographic groups
- **Case Studies**:
  - Impact of biased AI in healthcare, hiring, or law enforcement

Strategies for Fair AI Development and Bias Mitigation

- **Data-Related Strategies**:
  - Collect diverse and representative data
  - Use synthetic data to balance underrepresented groups
  - Perform bias audits on datasets
- **Algorithmic Strategies**:
  - Implement fairness-aware algorithms (e.g., adversarial debiasing)
  - Use regularization techniques to minimize bias
  - Evaluate models on fairness metrics
- **Operational Strategies**:
  - Foster diversity in AI teams
  - Continuous monitoring of AI systems in production
  - Transparent reporting of model behavior

Practice

Analysis of test cases to fultill the selected AI regulations. Discussion on ethics boundaries and balance between multiple factors

Užduotys

- Šiandien pradėsite daryti antrąjį namų darbą
- Dirbama komandomis po 5 – 6 žmones:
  - Tikslas turėti įvairaus profilio komandos narius
  - Kiekvienas Jūsų turite temą, komandoje visos šios temos bus aptariamos, diskutuojama ir randami tinkami šaltiniai ir reguliavimo priemonės
  - Remiatės 4 paskaitoje pateiktais dokumentais

Finansinės paslaugos

- **AI naudojimas:**
  - Finansų institucijos taiko DI kreditų reitingavimui ir sukčiavimo aptikimui
- **Reguliaciniai reikalavimai:**
  - Pavyzdžiui, JAV įmonės privalo laikytis **Fair Credit Reporting Act (FCRA)** nuostatų, kurios reikalauja, kad AI modeliai nebūtų diskriminaciniai
- **Testavimo scenarijai:**
  - Institucijos kuria testavimo atvejus, siekdamos įvertinti, ar DI sprendimai yra sąžiningi ir nešališki įvairioms demografinėms grupėms.
  - Tai apima skirtingų grupių rezultatų analizę, tikrinant, ar nėra nepagrįstų skirtumų tarp jų.

  - KAS YRA EUROPOJE?

Finansinės paslaugos

- Kreditingumo vertinimai. DI sistemos, vertinančios asmenų kredito balus ar mokumą, priskiriamos prie **didelės rizikos** kategorijos.
- Draudimo rizikos vertinimai. DI naudojamas gyvybės ir sveikatos draudimo kainų nustatymui bei rizikos vertinimui taip pat laikomas **didelės rizikos** pritaikymu
- Rizikos valdymas ir valdymo struktūra
- Duomenų kokybė ir skaidrumas
- Žmogiškos priežiūros mechanizmai

Sveikatos sektorius

- Rizikos valdymas ir valdymo struktūra
- Duomenų kokybė ir skaidrumas
- Žmogiškos priežiūros mechanizmai

- GDPR:
  - Specialios kategorjos duomenys
  - Teisėtas duomenų tvarkymas
  - Duomenų subjekto teisės

Įdarbinimas ir darbuotojų valdymas

- DI sistemos, naudojamos darbuotojų atrankai, veiklos vertinimui ar darbo jėgos valdymui, laikomos didelės rizikos
- Draudžiama naudoti DI emocijų atpažinimo sistemas darbo vietoje, nebent tai pateisinama medicininiais ar saugumo tikslais
- Gali būti renkama tik ta informacija, kuri yra būtina įdarbinimo ar darbo tikslams, ir ji negali būti naudojama kitiems tikslams

Autonominės TP

- Autonominių transporto priemonių (AV) veikimui esminiai DI komponentai laikomi didelės rizikos dėl jų tiesioginio poveikio keleivių saugumui ir visuomenės gerovei.
- Nustatykite ir klasifikuokite AV naudojamus DI komponentus, kad būtų aišku, kokie reglamentai taikomi.
- Būtinas reguliarus auditas

Žingsniai

- Reguliacinių reikalavimų identifikavimas
  - Suprasti kontekstą (koks sektorius, taikymo sritis)
  - Nustatyti reikalavimus pagal regioną
  - Sujunkite aktualius reglamentus
  - Atlikite vertinimą
  - Paruoškite atitikties strategiją
- Atitikties testavimo atvejų kūrimas

Žingsniai

- Šališkumo aptikimo testų kūrimas
    - Kokie tikslai?
    - Parinkite tinkamas metrikas
    - Parengti test cases
    - Automatizuoti šališkumo tikrinimą ir šalinimą
- Duomenų privatumo priemonių validavimas (Suprasti duomenų privatumo reikalavimus, užtikrinti duomenų tikrinimą, tikrinti)
- Paaiškinamumo ir skaidrumo vertinimas (paaiškinti kaip veikia modelis apskritai, ar įmanoma paaiškinti konkretų sprendimą, ar aišku iš kur duomenys, ar dokumentuota ir t.t.)
- Stebėsena ir nuolatinis testavimas

Pavyzdiniai šaltiniai:

- https://www.sciencedirect.com/science/article/pii/S1532046424000649
- https://www.sciencedirect.com/science/article/abs/pii/S0160791X24003087
- https://www.sciencedirect.com/science/article/abs/pii/S1547527124031102
- https://www.sciencedirect.com/science/article/pii/S2307187724001871
- https://www.sciencedirect.com/science/article/pii/S2214635024000868
- https://www.sciencedirect.com/science/article/pii/S2468227624002266
- https://www.sciencedirect.com/science/article/pii/S266682702400001X

Šaltiniai darbui

- https://aif360.res.ibm.com/
- ttps://github.com/fairlearn/fairlearn

Sources

- Tobias Baer. Understand, Manage, And Prevent Algorithmic Bias: A Guide For Business Users And Data Scientists
- Andyi Burkov. Machine Learning Engineering