2023 embedded VISION SUMMIT

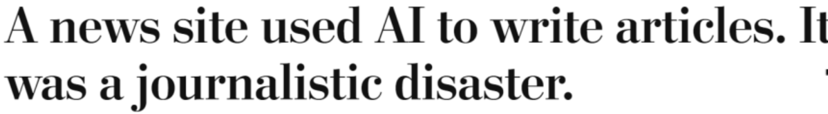
Responsible AI: Tools and Frameworks for Developing AI Solutions

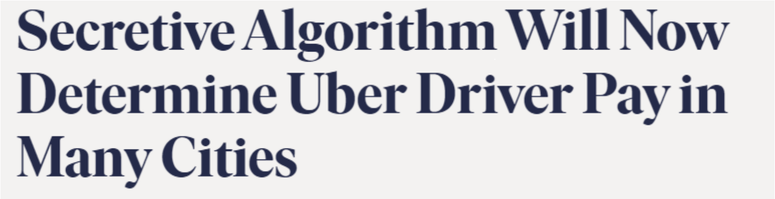
Mrinal Karvir

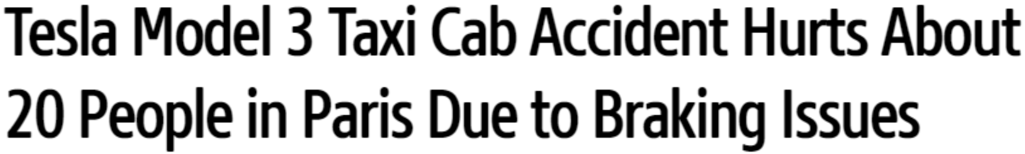
Senior Cloud Software Engineering Manager

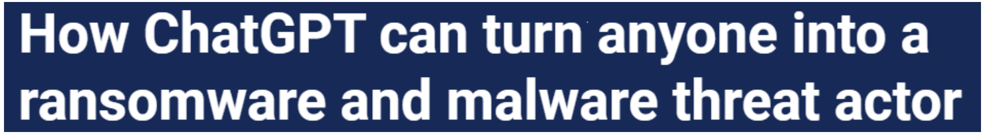
Intel Corporation

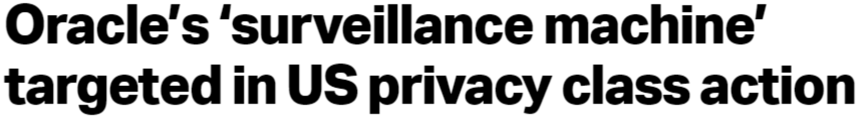
AI Incidents in the News

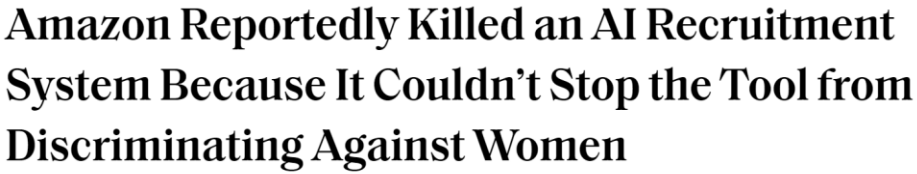

A news site used AI to write articles. It was a journalistic disaster.

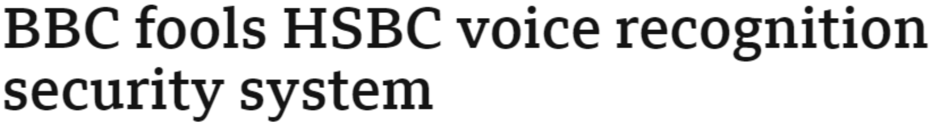

Secretive Algorithm Will Now Determine Uber Driver Pay in Many Cities


Tesla Model 3 Taxi Cab Accident Hurts About 20 People in Paris Due to Braking Issues


How ChatGPT can turn anyone into a ransomware and malware threat actor


Oracle’s ‘surveillance machine’ targeted in US privacy class action


Amazon Reportedly Killed an AI Recruitment System Because It Couldn’t Stop the Tool from Discriminating Against Women


BBC fools HSBC voice recognition security system

https://incidentdatabase.ai/

Cost of AI Incidents

Harm to human life

Loss of trust

Fines in compliance & regulations

Introduction of systemic bias

Misinformation

Breach of privacy

Defining Responsible AI Principles

Respect Human Rights

AI solutions should not support or tolerate usages that violate human rights.

Equity and Inclusion

Focus on data used for training and the algorithm development process to help prevent bias and discrimination.

Transparency

Understand and explain where the data came from and how the model works.

Enable Human Oversight

Human oversight of AI solutions to ensure they positively benefit society and do no harm.

Personal Privacy

Maintaining personal privacy and consent. Focusing on protecting the collected data.

Security, Safety, Sustainability

Ethical review and enforcement of end-to-end AI safety.
Low-resource implementation of AI algorithms

Designing with a Human Centric Approach

Definition

Does AI add value?

Who are the indented users of the system?

Identify unintended potential harm and plan for remediations

Translate user needs into data needs

Development

Source high-quality unbiased data responsibly

Get inputs from domain experts

Enable human oversight

Built-in safety measures

Deployment

Provide ways for users to challenge the outcome

Provide manual controls when AI fails

Offer high-touch customer support

Marketing

Focus on the benefit, not the technology

Transparently share the limitations of the system with the users

Be transparent about privacy and data settings

Anchor on familiarity

Bias Identification

- Historical bias
  - Can arise even if data is perfectly measured and sampled
  - The world as it is or was leads to a model that produces harmful outcomes
- Representation bias
  - When defining the target population, if it does not reflect the use population
  - When defining the target population, if contains underrepresented groups
  - When sampling from the target population, if the sampling method is limited or uneven
- Measurement bias
  - The proxy is an oversimplification of a more complex construct
  - The method of measurement varies across groups
  - The accuracy of measurement varies across groups

Bias Identification

- Aggregation bias
  - Arises from a one-size-fits-all model for data with underlying groups that are different
  - Results in model not optimal for any group/ favors the dominant group
- Learning bias
  - Arises when modeling choices amplify performance disparities across different examples in the data
  - Arise when prioritizing one objective damages another
- Evaluation bias
  - The benchmark data used for a particular task does not represent the use population
- Deployment bias
  - Arises due to "off-the-label" usage of model other than intended use

Case Study

Amazon’s hiring algorithm – The company's experimental recruiting tool utilized artificial intelligence to assign job applicants ratings.

- **Historical bias:** As a result of analyzing resumes for a decade, Amazon's computer models can spot similarities in candidates' applications.
- **Representation bias:** Most were from males, reflecting the industry's male dominance. Amazon's algorithm learned that male applicants were preferred. So, it penalized resumes that indicated that the applicant was female.
- **Aggregation bias:** It also demoted applications of those who attended one of two all-female institutions.

Tools to Detect & Mitigate Bias

- **What-If Tool (Google)**
  - Simulation with data manipulation & specific criteria to detect bias
  - 5 fairness metrics
  - Bias mitigation not straightforward
  - Open-source toolkit
  - Rich visualization
  - Notebook, Google cloud, Tensorboard
  - Rich tutorials and documentation

- **AI Fairness 360 (IBM)**
  - Extensible toolkit for bias detection & Mitigation
  - 70+ fairness metrics
  - 10 bias mitigation algorithms
  - open-source toolkit
  - Fairness metric explanations
  - Notebook, Python API, R code
  - Rich tutorials and documentation

Improve Transparency with Model Cards

- Shared understanding of AI models – intended use, performance, limitations
- Automation for the generation of Model Cards
  - Model Card Toolkit (TensorFlow)
  - pip install model-card (Python)
  - Default cards with Metaflow
  - Build in-house

GPT-3 Model Card

Model Details

Model date

Model type

Model version

Paper & samples

(CONTENT WARNING: GPT-3 was trained on arbitrary data from the web, so samples may contain offensive content and language.)

Model Use

Data

Performance

Limitations

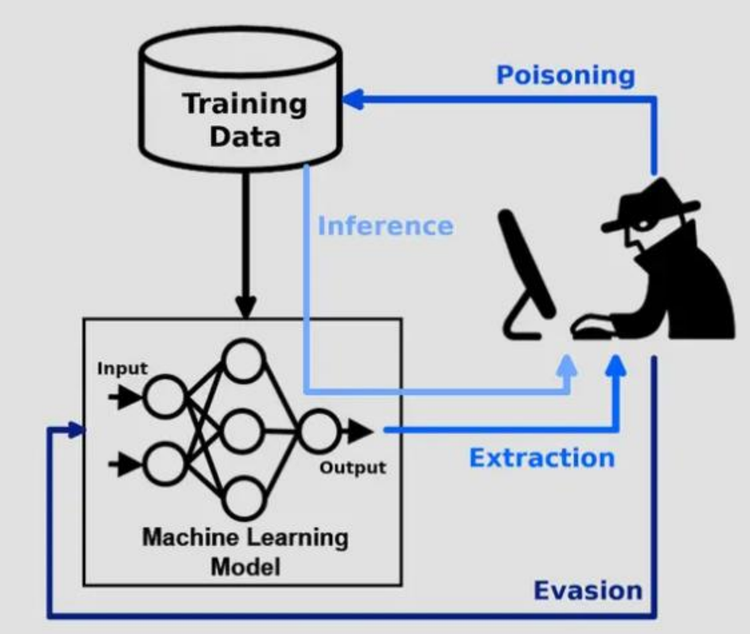
Where to send questions or comments about the model

Preserving Privacy

- Differential privacy - Publicly sharing information about a dataset while withholding information about individuals.
    - Open-source differential privacy library (Google) - Go, C++, and Java
    - PipelineDP (OpenMined , Google) - Python library
    - OpenDP (IQSS, Harvard, Alfred P. Sloan Foundation) - Rust, Python SDK with SQL Interface
  - Privacy vs accuracy trade-off
- Federated learning - Trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them. Requires enough local computing power and memory and high bandwidth connections
    - OpenFL (Intel Corporation)
    - TensorFlow Federated (Google)
    - PySyft (OpenMined)
    - IBM Federated Learning (IBM)

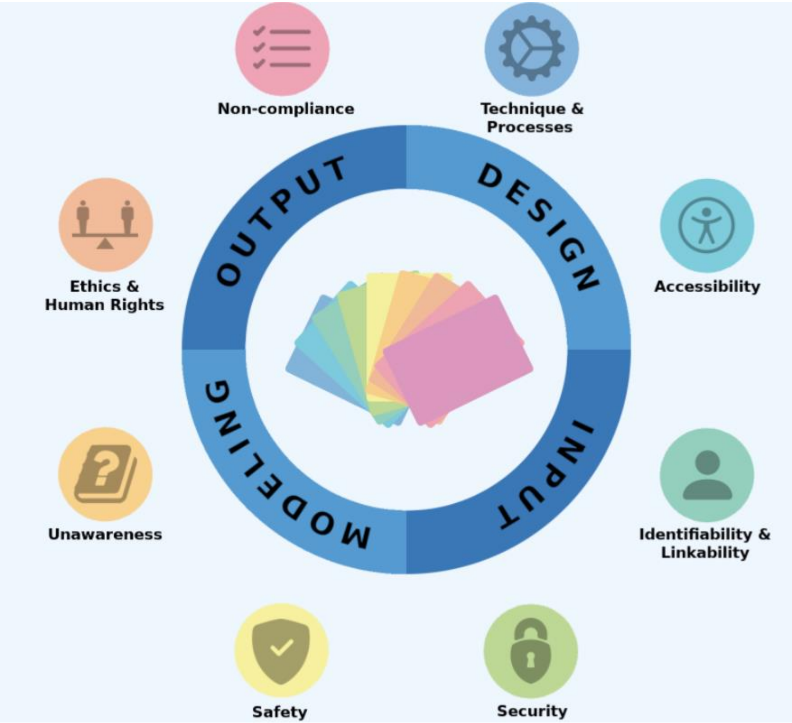
Security Toolkits

- Adversarial Robustness Toolbox (ART)
    - Python library for red & blue teams
    - Supports all popular machine learning frameworks
    - Data types supported - images, tables, audio, video
- Adversarial ML Threat Matrix - an industry-focused open framework, to empower security analysts to detect, respond to, and remediate threats against ML systems.



Threat Modeling for RAI with PLOT4AI

- PLOT4AI – Privacy by design
- Threat modeling library to help you build responsible AI
- 86 threats classified under the following 8 categories
- Provides threats in the form of card with questions to bring right focus & recommendations
- GDPR compliance focus



Responding to an AI incident

- Keeping up-to-date with the evolving legislature and regional laws - EU AI act, US federal AI regulations, US state AI regulations, China AI regulations and more
- Liability triage
- Identification/containment/recovery plan
- Communication plan
  - What happened?
  - Why did the incident happen?
  - What the entity has done or is going to do to?
  - Remediation and prevention plan
- Lessons learned

Microsoft Responding to its Offensive Chatbot

Learning from Tay's introduction

Mar 25, 2016 | Peter Lee - Corporate Vice President, Microsoft Healthcare

What happened?

As many of you know by now, on Wednesday we launched a chatbot called Tay. We are deeply sorry for the unintended offensive and hurtful tweets from Tay, which do not represent who we are or what we stand for, nor how we designed Tay.

Why did the incident happen?

The logical place for us to engage with a massive group of users was Twitter. Unfortunately, in the first 24 hours of coming online, a coordinated attack by a subset of people exploited a vulnerability in Tay. Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack. As a result, Tay tweeted wildly inappropriate and reprehensible words and images. We take full responsibility for not seeing this possibility ahead of time

Remediation and next steps

We will take this lesson forward as well as those from our experiences in China, Japan and the U.S. Right now, we are hard at work addressing the specific vulnerability that was exposed by the attack on Tay. Tay is now offline and we'll look to bring Tay back only when we are confident we can better anticipate malicious intent that conflicts with our principles and values.

Ethical AI Certifications

- **Why certification?**
  - Stay compliant with the latest AI regulations, be ready for emerging standards
  - Gain consumer, investor, regulator and insurer confidence on product’s safety, performance and reliability
- **Responsible Artificial Intelligence Institute (RAII)** -
  - First independent, accredited certification program for RAI
  - Vectors: Systems Operations, Explainability and Interpretability, Accountability, Consumer Protection, Bias and Fairness, and Robustness with collaborations across World Economic Forum, OECD, IEEE, ANSI, etc.,
  - Operates in the US, Canada, Europe and the United Kingdom
- **IEEE CertifAIEd™** -
  - Vectors: Ethical Privacy, Algorithmic Bias, Transparency, and Accountability.
  - Based on more than 25+ IEEE Standards on Ethically Aligned Autonomous and Intelligent Systems, Age Appropriate Design, Adaptive Instructional Systems and Machine Learning.

Key Take Aways

- Establish RAI principles that guide the decision-making for your AI development.
- Drive RAI requirements into product definition.
- Adopt a human-centric approach at every stage of your product development.
- Integrate RAI tools in your software development lifecycle.
- Preventing bias is complex. Define fairness metrics, document trade-offs and share with your users transparently. Re-check for bias often.
- Conduct regular assessments, audits, and update AI Response plans.
- Keep up-to-date with the evolving legislation, regional laws and standards.
- Certifications can help adherence to standards and legislation to build user trust.

Acknowledgements/Resources

- Suresh, Harini, and John Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle." In *Equity and access in algorithms, mechanisms, and optimization*, pp. 1-9. 2021
- Responsible AI Landscape : https://hai.stanford.edu/news/2022-ai-index-industrialization-ai-and-mounting-ethical-concerns (accessed in Feb 2023)
- https://pair.withgoogle.com/guidebook/ (accessed in Feb 2023)
- Tutorial: 21 fairness definitions and their politics - Arvind Narayanan
- GPT-3 Model Card: https://github.com/openai/gpt-3/blob/master/model-card.md (accessed in Feb 2023)
- PLOT4AI: https://plot4.ai/ (accessed in Feb 2023)
- Differential Privacy: https://en.wikipedia.org/wiki/Differential_privacy (accessed in Feb 2023)
- AI Incident Response Checklist: https://bnh-ai.github.io/resources/ (accessed in Feb 2023)
- Responsible Artificial Intelligence Institute Certification: https://www.responsible.ai/how-we-help (accessed in Feb 2023)
- IEEE CertifAIEd™ : https://engagestandards.ieee.org/ieeecertifaied.html (accessed in Feb 2023)

Backup

Common Fairness Metrics

At least 21 fairness metrics. Many are conflicting. Which is the fairest? No right answer. Some common metrics -

- Group Unaware - Removes all group and proxy-group membership information from the dataset. Difficult to achieve.
- Group Threshold - Alternate thought-process to group unaware. Optimize a separate threshold for each group for similar percentages of correct predictions
- Demographic parity - Similar percentages of datapoints from each group are predicted as positive classifications.
- Equal opportunity - Among those datapoints with the positive ground truth label, there is a similar percentage of positive predictions in each group.
- Equal accuracy - There is a similar percentage of correct predictions in each group