Trusted AI

Or how to build trustworthy machines

Patricia Ferreiro, IBM
@pferreiro92
Who am I?

Patricia Ferreiro

- Big Data & Analytics Architect @ IBM
- BSc Electrical Engineering
- Currently MSc Data Science
- Based in Barcelona, Spain
- Aspiring polyglot

@pferreiro92
Do we trust AI?
AI that detects cardiac arrests during emergency calls will be tested across Europe this summer

The software listens in to calls and helps emergency dispatchers make judgements

By James Vincent | Apr 25, 2018, 10:06am EDT
161,650 emergency calls related to cardiac arrests.

Source: European Emergency Number Association (EENA) and Corti
161,650 emergency calls related to cardiac arrests.

AI was more precise than human operators: 95.3% vs 73.9% detections...

Source: European Emergency Number Association (EENA) and Corti
161,650 emergency calls related to cardiac arrests.

AI was more precise than human operators: 95.3% vs 73.9% detections...

...and faster: 48 vs 79 seconds on average

Source: European Emergency Number Association (EENA) and Corti
Do we trust AI?

In traditional software development, trust is built through standardized processes such as testing suites, audit procedures or documentation.

However, AI systems build knowledge up over time, are non-deterministic and often difficult to understand.
AI adoption by high-stakes decision making applications is increasing exponentially. Nowadays, AI helps answer many questions:

- Which resumes are considered?
- What is your medical diagnosis?
- Who gets their mortgage loan approved?
- Will your car stop to avoid danger?
- ...
The four pillars of Trusted AI

- **Fairness**: Is it ethical?
- **Explainability**: Is it easy to understand?
- **Robustness**: Is it reliable?
- **Lineage**: Is it accountable?

@pferreiro92
Fairness
Fairness

Motivation

Source: COMPASS Software Results’, Julia Angwin et al. ‘16
Fairness

Motivation II

• The **quality** of an AI system is as good as the data it feeds on

• AI should not learn and propagate our biases

• To create fair applications we must detect and mitigate bias throughout the lifecycle of AI systems

@pferreiro92
Fairness

Motivation III

Biased AI systems can have a negative impact on critical areas:

- **Product usability**
- **Laws and regulations**
- **Ethical issues**
Can you quantify the impact your AI models have in your business as well as in your client opportunities?

@pferreiro92
How is bias measured?

- Multiple definitions of bias exist
  - Statistical measures
  - Similarity-based measures
  - Causal reasoning

- Can be contradictory!

- Domain knowledge may be required

- Accuracy vs utility trade-off
Where does bias come from?

- Data acquisition/sampling
- Human labeling
- Propagated historical bias
- Algorithm design
- ...
Fairness

2018 Timeline

- Announces internal tool “Fairness Flow” now jointly developed with TU München
- Announces development of internal tools to evaluate bias
- Publishes “What-If tool”, a visual exploration tool including bias mitigation algorithms
- Publishes “AIF360” framework, with 30+ metrics, 9+ mitigation algorithms and a certain degree of explainability

@pferreiro92
Open Source tools

- **FairSearch**: [https://github.com/fair-search](https://github.com/fair-search)
  Framework for specific algorithm testing on multiple datasets and fairness measures.

- **FairML**: [https://github.com/adebayoj/fairml](https://github.com/adebayoj/fairml)
  Features four input ranking algorithms to quantify a model’s relative predictive dependence on model’s inputs.

- **FairTest**: [https://github.com/columbia/fairtest](https://github.com/columbia/fairtest)
  Learns a decision tree that splits a user population into smaller subgroups in which the association between protected features and algorithm outputs is maximized.

@pferreiro92
Open Source tools II

- **UChigago Aequitas**: [https://github.com/dssg/aequitas](https://github.com/dssg/aequitas)
  Produces a report on multiple statistical bias metrics.

- **PyMetrics Audit-AI**: [https://github.com/pymetrics/audit-ai](https://github.com/pymetrics/audit-ai)
  Built on top of pandas and sklearn, implements fairness-aware ML algorithms with metrics for both classification and regression tasks.
Fairness

Open Source tools III

- IBM AIFairness360: [https://github.com/IBM/AIF360](https://github.com/IBM/AIF360)
  Framework for bias statistical assessment and mitigation throughout the model lifecycle.

@pferreiro92
Fairness

Lessons learned

- Bias appears in the data and **may inaccurately model populations**
- Mitigating bias **may decrease model accuracy**
- Bias assessment and mitigation is an iterative and complex process
  - Mostly not regulated
  - Fuzzy domain-specific definitions
- **Several open source initiatives 😊**

@pferreiro92
Explainability

Motivation

Source: Clinically applicable deep learning for diagnosis and referral in retinal disease, De Fauw et al. 2018
Explainability

Motivation II

• Understanding **how AI systems arrive at an outcome** is key to trust

• Humans are **legally and morally liable**

• To improve transparency, **local and global interpretability** of AI models is required
The European Union’s General Data Protection Regulation (GDPR) grants consumers the right to know when automated decisions are being made about them and the right to have these decisions explained.

Enterprises that adopt XAI now will be prepared for future compliance mandates.

Source: https://gdpr-info.eu/art-22-gdpr/
Explainability

Accuracy vs Explainability trade-off

Source: DARPA (US Department of Defense) XAI Project
Explainability

Technical approaches

• **Explanation by Design**

• **Black Box eXplanation**
  1. Train a complex model on some dataset
  2. Train an interpretable model on the original dataset plus the predictions

@pferreiro92
Explainability

Black Box eXplanation

Local Interpretable Model-agnostic Explanations - LIME


@pferreiro92
Black Box eXplanation II

SHapley Additive exPlanations - SHAP

Source: A Unified Approach to Interpreting Model Predictions, Scott M. Lundberg et al. 2017

@pferreiro92
Black Box eXplanation III

Neural nets – Feature visualization

Source: Feature visualization, Google Brain ‘17
Explainability

Black Box eXplanation IV

Neural nets – Feature visualization

Source: Feature visualization, Google Brain ‘17
Black Box eXplanation V

Source: Learning Deep Features for Discriminative Localization, MIT ‘15

@pferreiro92
Explainability

Open Source tools

- **LIME**: [https://github.com/marcotcr/lime](https://github.com/marcotcr/lime)

  Supports local explainability for images, text classifiers and classifiers that act on tables. Visualizations are generated in HTML and matplotlib.
Open Source tools II

- **SHAP:** [https://github.com/slundberg/shap](https://github.com/slundberg/shap)

  Provides explainers for any ML model by generalizing multiple Additive Feature Attribution Methods such as LIME, connecting game theory with a local explanation. Generates JS visualizations. **SHAP values represent a feature's responsibility for a change in the output.**
Explainability

Open Source tools II

- **SHAP**: [https://github.com/slundberg/shap](https://github.com/slundberg/shap)
• **Google What-If Tool:**

  TF plugin for visually investigating model performance and fairness over subsets of a dataset and counterfactual exploration.
Explainability

Lessons learned

• Transparency is key for “Augmented AI” to be widely adopted

• Explainability must be taken into account during algorithm design

• Powerful, extensible open source frameworks for generating explainable models already exist
Robustness
Robustness

Motivation

Logo Attacks

Original

Adversarial

Classified as: Stop No overtaking

Custom Sign Attacks

Speed limit (30) Stop

Adversarial Traffic Signs

Original

Adversarial

Classified as: Stop Speed limit (30)

Source: Deceiving Autonomous Cars with Toxic Signs, Princeton University

@pferreiro92
Robustness

Motivation II

- AI systems aim to act autonomously in critical scenarios where a single mistake may have a high cost

- State of the art AI systems have been proven weak against relatively simple adversarial inputs
Robustness

Defining robustness against...

- **Human errors**: type check, variable ranges
- **Malicious attacks**: adversarial inputs
- **Incorrect models**: regularization, risk-sensitive objectives
- **Unmodeled phenomena**: expand model
  - It is impossible to model everything
  - It is not desirable to model everything

*AI systems must be able to act autonomously without having a complete model of the world*
Robustness

Key insights

• **Minimal perturbations**, often imperceptible to humans, that completely fool AI systems into unwanted behaviour (2013, Szegedy et al.)

• A practical definition of the robustness of a model is the average size of the minimum adversarial perturbation.

• **Black vs White box** attacks: on training or serving step.
• **One-time vs Iterative** attacks: one-time are highly transferrable and thus more effective in black box attacks.
Robustness

Types of adversarial attacks

• **Gradient-based**: finds directions to which the model predictions for a given class are most sensitive to.

• **Score-based**: use class probabilities or logits to approximate gradients.

• **Decision-based**: rely only on the class decision of the model.
Robustness

Adversarial attacks I

Source: Accessorize to a Crime, Mahmood Sharif et al., 2016
Adversarial attacks II

Source: Crafting Adversarial Examples For Speech Paralinguistics Applications, Yuan Gong et al., 2017
Adversarial defenses

Adversarial images - JPEG Compression

Source: Defending AI with JPEG Compression, Nilaksh Das, ‘17
Robustness

Open Source tools

• **Borealis AI – AdverTorch**: [https://github.com/BorealisAI/advertorch](https://github.com/BorealisAI/advertorch)
  Attack and defense API for PyTorch.

• **Cleverhans**: [https://github.com/tensorflow/cleverhans](https://github.com/tensorflow/cleverhans)
  Benchmark AI systems vulnerability to adversarial examples. Roadmap: support for JAX, PyTorch, and TF2.

• **Foolbox**: [https://github.com/bethgelab/foolbox](https://github.com/bethgelab/foolbox)
  Extensible framework for adversarial robustness benchmarking, both implementing gradient-based attacks and black-box attacks. Supports multiple frameworks.
Robustness

Open Source tools II

- **IBM ART**: [https://github.com/IBM/adversarial-robustness-toolbox](https://github.com/IBM/adversarial-robustness-toolbox)
  
  Python library that implements adversarial attacks, defenses and robustness metrics for multiple ML and DL algorithms with multiple framework support.

@pferreiro92
Robustness

Lessons learned

- AI systems are **not robust by default**
- **Testing and debugging practices** have not been standardized for AI
- Adversarial evaluation provides robustness metrics related to **model quality and security**

@pferreiro92
Lineage
In order to enable an AI marketplace each digital asset must be trackable, verifiable and held accountable
The dataset nutrition label

Source: The Dataset nutrition label - MIT, Harvard ’18
The dataset nutrition label

- Common metadata
- Provenance
- Variable description and statistics
- Pair plots
- Probabilistic models
- Ground truth correlations

Source: The Dataset nutrition label - MIT, Harvard ’18
IBM proposes a **Supplier’s Declaration of Conformity** (SDoC) that helps provide information about the **four key pillars of trusted AI**.

- Dataset “nutritional label”
- Bias assessment and mitigation
- Algorithm explainability and interpretability
- Robustness policy
Proposal overview

Source: An End-To-End Machine Learning Pipeline That Ensures Fairness Policies, IBM Research ‘17
AI is experiencing a renaissance and, according to Gartner, it’s vital that we

“build AI right, use AI right, keep AI right”.

The values adopted to build today’s AI systems will be reflected in the decisions those systems make for a decade or more.
Thank you! Q&A

@pferreiro92