Course 4: Practical Ethics in AI

Nicolas Farrugia

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Summary

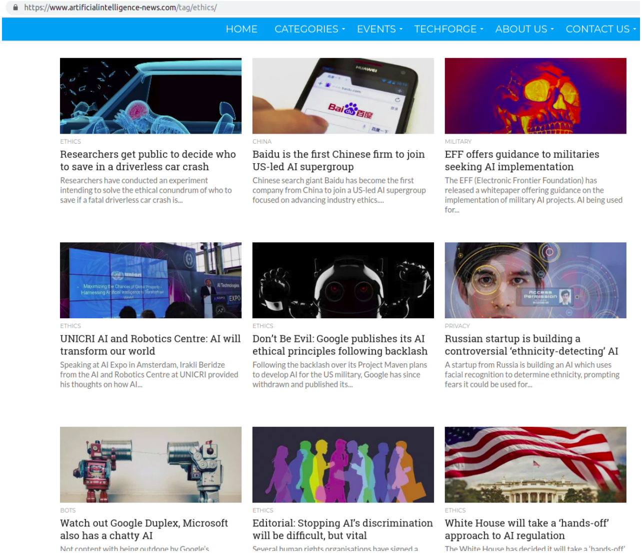
Last sessions

1. Supervised learning - learning from labeled examples
2. Unsupervised learning - discovering structure in data

Today's session

1. Generalities on Ethics in AI
2. Practical challenges in machine learning with ethical consequences

Why ?



https://www.artificialintelligence-news.com/tag/ethics/

ETHICS

Researchers get public to decide who to save in a driverless car crash

Researchers have conducted an experiment intending to solve the ethical conundrum of who to save if a fatal driverless car crash is...

CHINA

Baidu is the first Chinese firm to join US-led AI supergroup

Chinese search giant Baidu has become the first company from China to join a US-led AI supergroup focused on advancing industry ethics....

MILITARY

EFF offers guidance to militaries seeking AI implementation

The EFF (Electronic Frontier Foundation) has released a whitepaper offering guidance on the implementation of military AI projects. AI being used for...

ETHICS

UNICRI AI and Robotics Centre: AI will transform our world

Speaking at AI Expo in Amsterdam, Irakli Beridze from the AI and Robotics Centre at UNICRI provided his thoughts on how AI...

ETHICS

Don't Be Evil: Google publishes its AI ethical principles following backlash

Following the backlash over its Project Maven plans to develop AI for the US military, Google has since withdrawn and published its...

PRIVACY

Russian startup is building a controversial 'ethnicity-detecting' AI

A startup from Russia is building an AI which uses facial recognition to determine ethnicity, prompting fears it could be used for...

BOTS

Watch out Google Duplex, Microsoft also has a chatty AI

Not content with being outdone by Google's

ETHICS

Editorial: Stopping AI's discrimination will be difficult, but vital

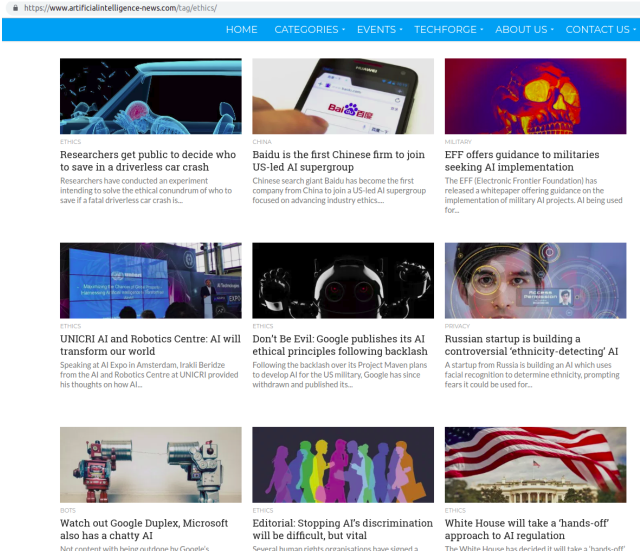
Several human rights organisations have signed a

ETHICS

White House will take a 'hands-off' approach to AI regulation

The White House has decided it will take a 'hands-off'

Why ?



ETHICS

Researchers get public to decide who to save in a driverless car crash

Researchers have conducted an experiment intending to solve the ethical conundrum of who to save if a fatal driverless car crash is...

CHINA

Baidu is the first Chinese firm to join US-led AI supergroup

Chinese search giant Baidu has become the first company from China to join a US-led AI supergroup focused on advancing industry ethics....

MILITARY

EFF offers guidance to militaries seeking AI implementation

The EFF (Electronic Frontier Foundation) has released a whitepaper offering guidance on the implementation of military AI projects. AI being used for...

ETHICS

UNICRI AI and Robotics Centre: AI will transform our world

Speaking at AI Expo in Amsterdam, Irakli Beridze from the AI and Robotics Centre at UNICRI provided his thoughts on how AI...

ETHICS

Don't Be Evil: Google publishes its AI ethical principles following backlash

Following the backlash over its Project Maven plans to develop AI for the US military, Google has since withdrawn and published its...

PRIVACY

Russian startup is building a controversial 'ethnicity-detecting' AI

A startup from Russia is building an AI which uses facial recognition to determine ethnicity, prompting fears it could be used for...

BOTS

Watch out Google Duplex, Microsoft also has a chatty AI

Not content with being outdone by Google's

ETHICS

Editorial: Stopping AI's discrimination will be difficult, but vital

Several human rights organisations have signed a

ETHICS

White House will take a 'hands-off' approach to AI regulation

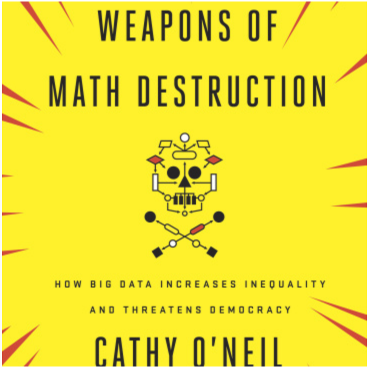
The White House has decided it will take a 'hands-off'

Our goal: distinguishing the hype from important developments.

Acknowledgment

This course is highly inspired from recommendations in the Villani report on AI (openly accessible), as well as O'neil's book.

CÉDRIC VILLANI
Mathématicien et député de l'Essonne
DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE
POUR UNE STRATÉGIE NATIONALE ET EUROPÉENNE


WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL

Five Challenges relating Ethics and AI

Regulatory and societal aspects

- Collective rights regarding data
- Keeping control on what (not) to develop
- A specific governance for Ethics in AI and public debate

Technical aspects

- Black-Boxes, transparency and bias
- Integrating ethics in engineering / design

Regulatory and societal aspects

Collective rights regarding data

- Existing regulations on (individual) private data (e.g. GDPR)
- No common policies on collective rights - group data

Keeping control

- Open solutions for auditing / controlling
- Non-proliferation of autonomous weapons

A specific governance for Ethics in AI

- Towards specific governance (consulting councils?)
- Role of public debate and transparency

Regulatory and societal aspects

Collective rights regarding data

- Existing regulations on (individual) private data (e.g. GDPR)
- No common policies on collective rights - group data

Keeping control

- Open solutions for auditing / controlling
- Non-proliferation of autonomous weapons

A specific governance for Ethics in AI

- Towards specific governance (consulting councils?)
- Role of public debate and transparency

Regulatory and societal aspects

Collective rights regarding data

- Existing regulations on (individual) private data (e.g. GDPR)
- No common policies on collective rights - group data

Keeping control

- Open solutions for auditing / controlling
- Non-proliferation of autonomous weapons

A specific governance for Ethics in AI

- Towards specific governance (consulting councils?)
- Role of public debate and transparency

Black-Boxes, transparency and bias 1/2

The problem of black boxes

- Trust by users
- Verifiability

Bias

- Reproducing the biases seen in society
- Potentially difficult to detect

Black-Boxes, transparency and bias 1/2

The problem of black boxes

- Trust by users
- Verifiability

Bias

- Reproducing the biases seen in society
- Potentially difficult to detect

Black-Boxes, transparency and bias 2/2

Tackling interpretability

Neural networks, Random Forest (and others) are difficult to interpret.

- Interpretability is an active research field,
- Procedures to explain algorithms by manipulating data.

Auditing AIs ?

Trust in AI approaches can potentially be increased using:

- Open-source and open data,
- Specific test procedures targetted to "fool" algorithms, to evaluate their robustness.

Integrating ethics in engineering / design

Dataset construction

Not always trivial to collect data...

- Because humans collect data, data can reproduce human biases.
- In some cases, exceptions, irregularities and accidents are more significant than the norm.

Training and benchmarking

It is essential to systematically consider:

- Accuracy, precision and recall
- Cross-validation

Some examples

- Open AI develops all-open solutions for AI, including code and data.
- Facebook AI Research publishes only open access papers and publishes all associated code.
- Google Open-sourcing some of its software.
- Parcours Sup's algorithm is open-source.

OpenAI

Facebook AI Research

Google AI

parcoursup
Entrez dans l'enseignement supérieur

Lab Session 4 and assignments for Session 5

Unsupervised learning Lab session

Finish the lab session on unsupervised learning (->9h30)

Lab session on practical aspects related to ethics

- Low sample size problem
- Imbalanced datasets

Project 2 (P2)

You have chosen an unsupervised learning method. You have to prepare a Jupyter Notebook on this method, including:

- A brief description of the theory behind the method,
- Advanced tests and analysis on your own PyRat Datasets.

During Session 5 (November 21st) you will have 7 minutes to present your notebook.