Practical Ethics in Artificial Intelligence

Nicolas Farrugia

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

Summary

Other sessions

1. Supervised learning - learning from labeled examples
2. Unsupervised learning - discovering structure in data
3. Reinforcement Learning - learning how to get better from reward
4. Combinatorial Game Theory - exploring various solutions to a problem

Today's session

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

Why ?

Search AI and Ethics ?

Why ?

1. Hype vs true risks, and associated Technical Challenges.
2. Technical Challenges can become ethical issues:
   - Dataset biases (lack of diversity)
   - Overfitting
   - Imbalanced classes
   - Reward definition
   - ...

Acknowledgment

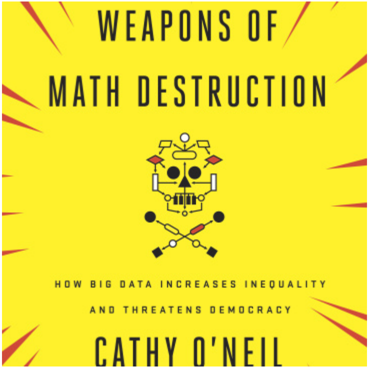
This course is highly inspired from recommendations in the Villani report on AI (openly accessible), as well as O'neil's book.

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

Also another recent good read :
`https://www.journalofdemocracy.org/ai-and-catastrophic-risk/`

Technical Challenges relating Ethics and AI

Regulatory and societal aspects

- Collective rights regarding data
- Keeping control on what (not) to develop
- Governance

Technical aspects

- Black-Boxes, transparency and bias
- Integrating ethics in engineering / design
- Differential privacy
- Federated learning

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

Main issue: (statistical / data) relationship between single individuals and grouped data.

Keeping control

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

A similar issue than with nuclear weapons.

Regulatory and societal aspects

A specific governance for Ethics in AI

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

What can we do ?

Institutional proposals

- GDPR
- European union AI Act
- UNESCO Recommendation on the Ethics of Artificial Intelligence
- Montreal declaration

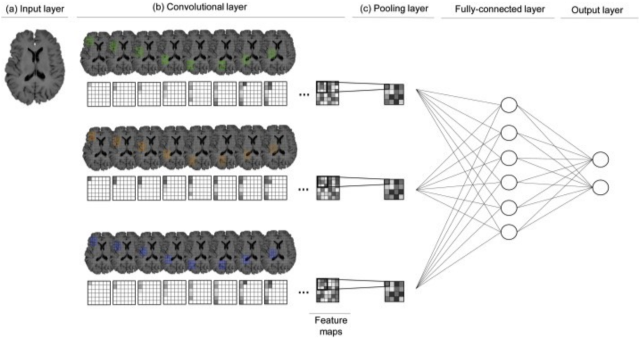
Technical aspects

- Black-Boxes, transparency and bias
- Integrating ethics in engineering / design
- Federated learning
- Differential Privacy

Black-Boxes, transparency and bias 1/2

The problem of black boxes

- Trust by users
- Verifiability



Black-Boxes, transparency and bias 1/2

Bias

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

Related technical problems in machine learning

- Difficulty to generalise from train to test due to a lack of diversity
- Similarity between train and test data
- Imbalanced classes

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 used to develop all-open solutions for AI…
- Facebook AI Research publishes only open access papers and publishes all associated code.
- Google Open-sourcing some of its software.

See the additional file with the list of ressources.