Practical Ethics in Artificial Intelligence

Nicolas Farrugia

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

Summary

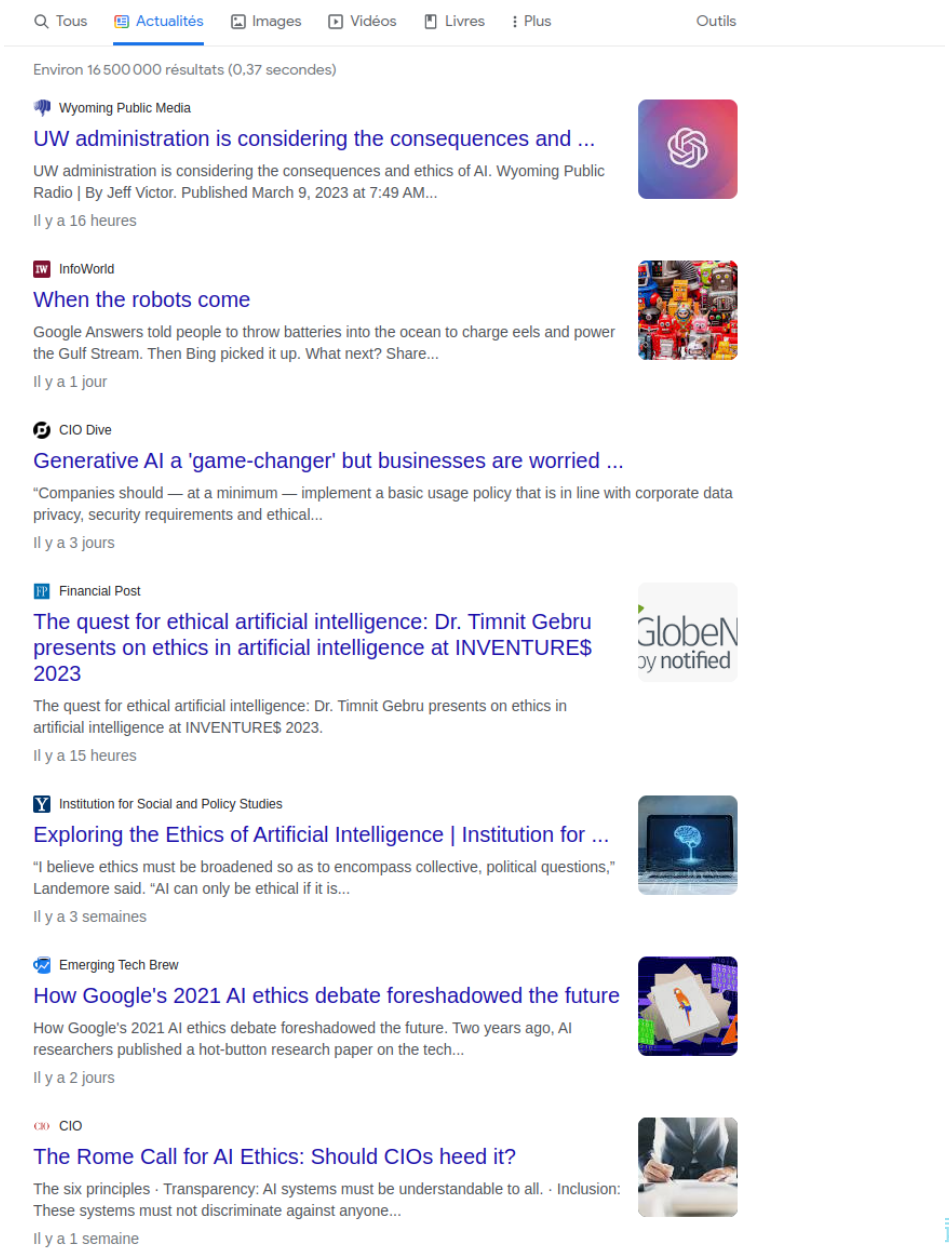
Other sessions

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 ?

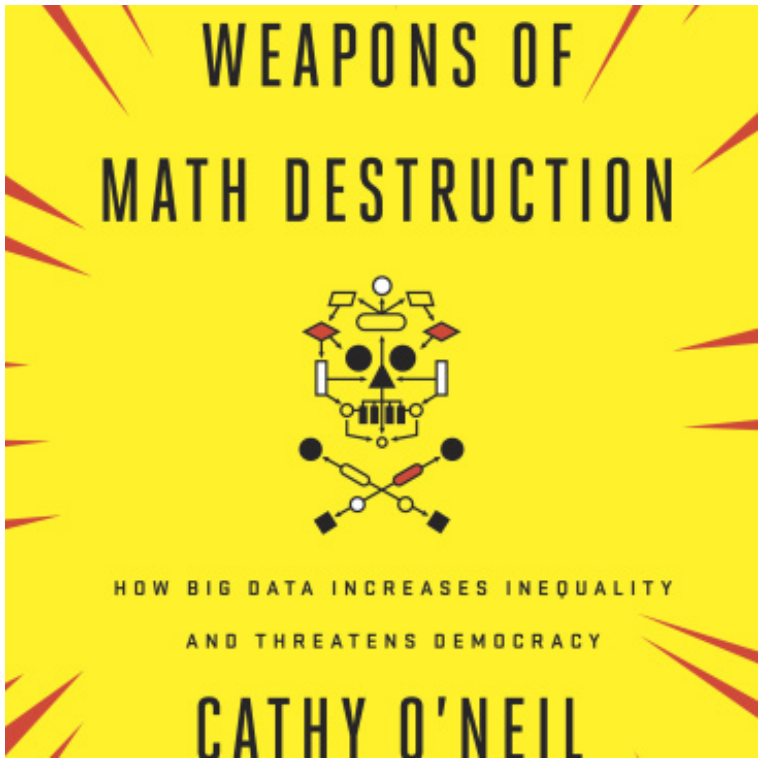

Tous
Actualités
Images
Vidéos
Livres
Plus
Outils
Environ 16 500 000 résultats (0,37 secondes)
Wyoming Public Media
UW administration is considering the consequences and ...
UW administration is considering the consequences and ethics of AI. Wyoming Public Radio | By Jeff Victor. Published March 9, 2023 at 7:49 AM...
Il y a 16 heures
InfoWorld
When the robots come
Google Answers told people to throw batteries into the ocean to charge eels and power the Gulf Stream. Then Bing picked it up. What next? Share...
Il y a 1 jour
CIO Dive
Generative AI a 'game-changer' but businesses are worried ...
"Companies should — at a minimum — implement a basic usage policy that is in line with corporate data privacy, security requirements and ethical...
Il y a 3 jours
Financial Post
The quest for ethical artificial intelligence: Dr. Timnit Gebru presents on ethics in artificial intelligence at INVENTURE$ 2023
The quest for ethical artificial intelligence: Dr. Timnit Gebru presents on ethics in artificial intelligence at INVENTURE$ 2023.
Il y a 15 heures
Institution for Social and Policy Studies
Exploring the Ethics of Artificial Intelligence | Institution for ...
"I believe ethics must be broadened so as to encompass collective, political questions," Landemore said. "AI can only be ethical if it is...
Il y a 3 semaines
Emerging Tech Brew
How Google's 2021 AI ethics debate foreshadowed the future
How Google's 2021 AI ethics debate foreshadowed the future. Two years ago, AI researchers published a hot-button research paper on the tech...
Il y a 2 jours
CIO
The Rome Call for AI Ethics: Should CIOs heed it?
The six principles · Transparency: AI systems must be understandable to all. · Inclusion: These systems must not discriminate against anyone...
Il y a 1 semaine

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.



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

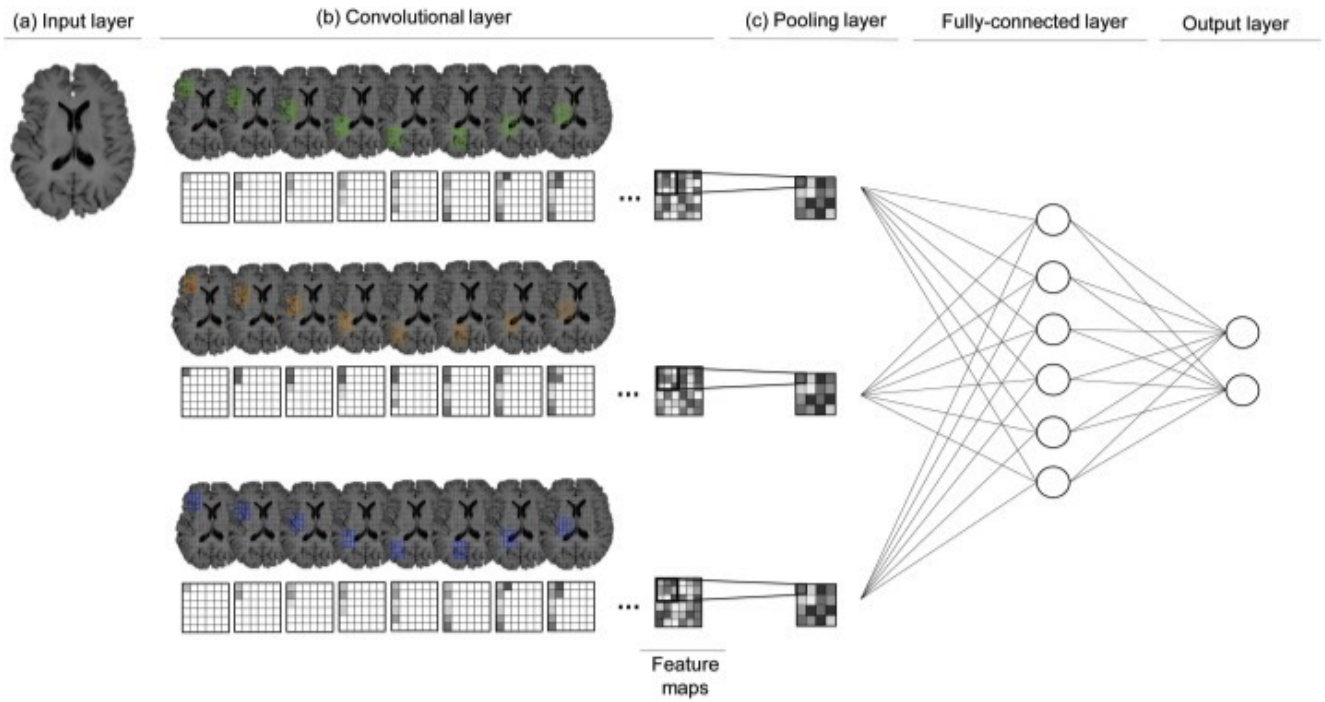
Technical aspects

- 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.