Goal, Risks and Countermeasures in the Artificial Intelligence Era

DataScienceSeed #11

Gabriele Graffieti

PhD Student @ University of Bologna
Head of AI Research @ AI for People

January 23, 2020

About me I

- PhD Student in Data Science and Computation @ University of Bologna.
- Head of AI Research @ AI for People.
- Teaching Assistant of the course Algorithms and Data Structures @ Unibo.
- Board Member of Data Science Bologna.
- Proud Member of *ContinualAI*.
- Amateur Astronomer @ Associazione Astrofili Cesenati

About me II

My main interests

- Generative models (GANs, VAEs, etc. . .)
- Continual/Lifelong Learning.
- Bio-inspired models.
- AI ethics and philosophy.
- The future evolution of AI.
- The social impact of AI (AI for social good).

AI for People

AI for People is an open community of people with different and heterogeneous backgrounds, with the goal of:

> *shaping Artificial Intelligent technology around human and societal needs. We believe that technology should respect the anthropocentric principle. It should be at the service of people, not vice-versa. In order to foster this idea, we need to narrow the gap between civil society and technical experts. This gap is one in knowledge, in action and in tools for change.*

AI for People

- Our mission is to learn, pose questions and take initiative on how Artificial Intelligent technology can be used for the social good.
- Our strategy is to conduct impact analysis, projects and democratic policies that act at the crossing of Artificial Intelligence and society.
- Learn more about *AI for People* at `https://www.aiforpeople.org`
- We always search for new people who want to get involved! Join us on slack here!

Why Ethics is Important

https://github.com/ggraffieti/evolutionary-cars

1. AI Background
2. Main Issues of AI
3. How to build ethical machines
4. AI for social good
5. AI for People

What is AI?

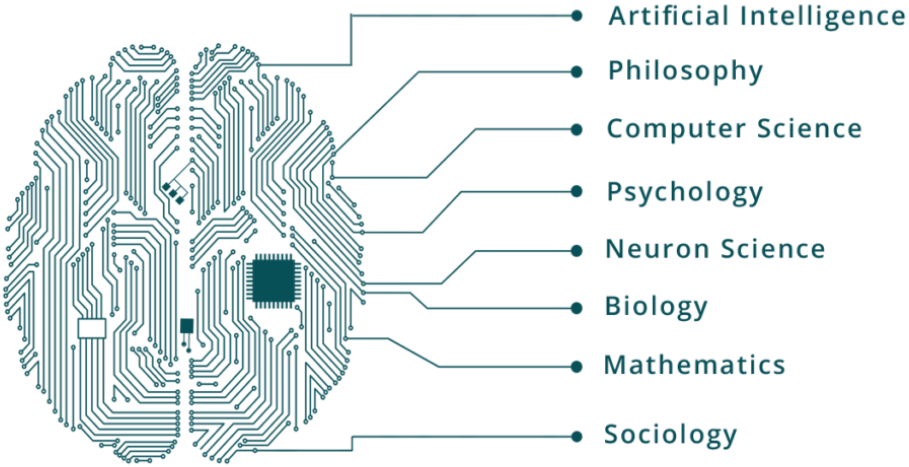
A branch of computer science dealing with the simulation of intelligent behavior in computers.

Merriam-Webster

The science and engineering of making intelligent machines.

John McCarthy

AI is a Broad Discipline I



AI is a Broad Discipline II

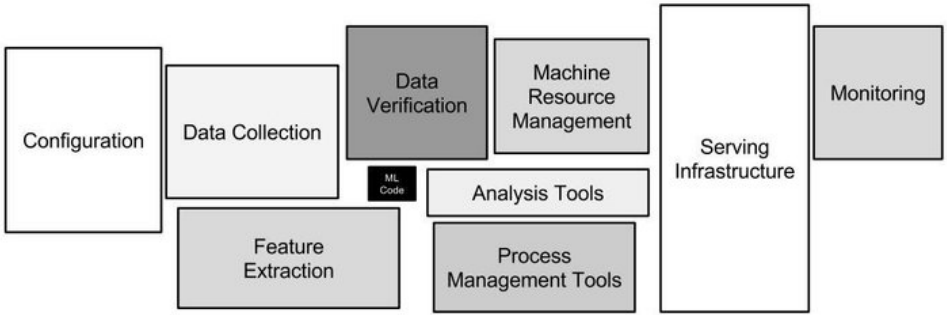


Image from *"Hidden Technical Debt in Machine Learning Systems"*, *Sculley et al.*

What Artificial Intelligence Can Do (Better than You) I

Computer Vision

- Image Classification, Segmentation, Captioning, ...
- Analysis of medical images, such as radiography, tomography, MR, ...
- Image Reconstruction (Denoising, Image completion, Super-resolution, ...)

Forecasting

- Trading, stock market prediction
- Traffic, people you may know, advertising, ...

What Machine Learning Can Do (Better than You) II

And many other things

- Speech recognition, translation, handwriting recognition
- Sentiment analysis, health monitoring, fraud detection, marketing
- DNA sequence classification, drug discovery, material discovery, theorem proving
- Resource allocation, robot locomotion, play games (chess, go, starcraft)
- And many many more things!

AI Strategies Around the World

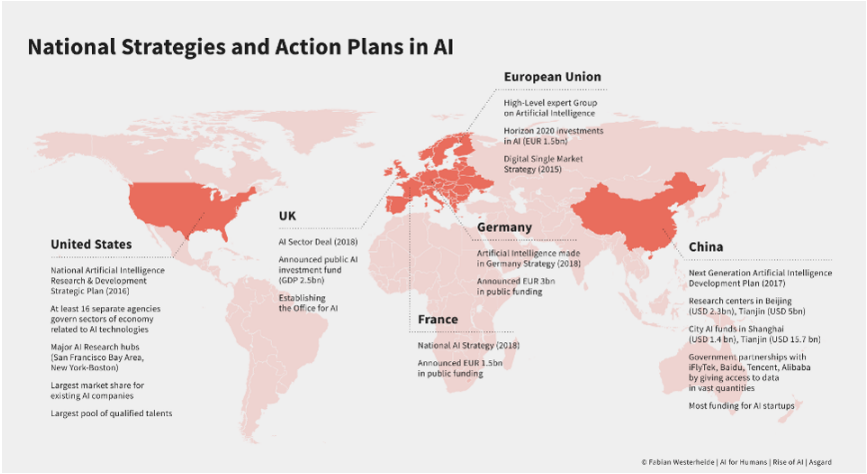


Image from *"The Artificial Intelligence Industry Global Challenges"*, *Forbes*.

EU High Level Expert Group on AI

"In my first 100 days in office, I will put forward legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence."

- Ursula von der Leyen,
President of the European Commission

Main Issues of AI

Who Owns AI?

AI needs (a big) infrastructure

- The algorithm is just a small part of the product.
- Computational capabilities (computational power and memory) are fundamental.
- Only the biggest companies have the workforce to maintain a solid infrastructure.
  $\rightarrow$ Substantial advantage over smaller companies or academia.

AI needs (a lot of) data

- Data is essential to reproduce results.
- Data is often more important than algorithm (who owns data?)
- Big tech companies have the possibility to acquire a huge amount of data daily.
  $\rightarrow$ Substantial advantage over smaller companies or academia.

The Myth of AI Democratization I

AI big companies claim to be democratic

- Sharing their research (e.g. arXiv).
- Sharing their code (e.g. github).
- Sharing their frameworks (e.g. Tensorflow).
- Sharing their infrastructure (?) (e.g. colab).

Technology democratization

[...] at an increasing scale, consumers have greater access to use and purchase technologically sophisticated products, as well as to participate meaningfully in the development of these products.

The Myth of AI Democratization II

AI is currently owned by few companies

- They have access to a huge amount of data.
- They attract top AI scientists (huge salaries, freedom).
- They have the power to transform research ideas into products.

Why AI democracy is important

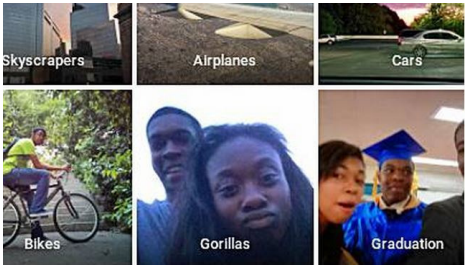
- Avoid monopolies.
- Openness in AI development could lower the probabilities of an AI singleton (Nick Bostrom).
- Democratization means that everyone gets the opportunities and **benefits** of artificial intelligence.

Data Bias I

AI, data or human bias?

When trained on man-made data, machine learning is likely to pick up the same constitutional and unconscious biases already present in society.

- SIRI was not able to recognize weird-accent speakers.
- Google face recognition tagged some black people as gorillas.



Data Bias II

COMPAS

COMPAS is a case management and decision support tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.

- The model and the approach was extremely ordinary.
- The data used was real and verified.
- The bias was not in model or data, but in the domain itself!

Food for thought $\rightarrow$ if a domain is intrinsically biased, the resulting model can be considered biased?

The Role and the Politics of Datasets I

Must read → `https://www.excavating.ai`

- Every label is a relation between the visual content of an image and a "meaning".
- In dataset meaning is assigned
- The set of "meanings" is flattened (e.g. bike and health).
- The relation between images and labels is build in an uncontrolled environment (e.g. Amazon Mechanical Turk).
- Food for thought → the concept of happiness can be described by an image?

The Role and the Politics of Datasets II

In ImageNet:

- There are categories such as: Bad Person, Call Girl, Drug Addict, Closet Queen, Convict, Crazy, Failure, Flop, ...
- The images are downloaded from the web without asking.
- People have been labeled without any permission.

← The woman here is labeled as "snob".

The Role and the Politics of Datasets III

Removing (or correcting) the dataset is not the solution!

- Until their data is released, it is impossible to do forensic testing on how they classify and interpret human bodies, actions, or inactions.
- If they are, or were, being used in systems that play a role in everyday life, it is important to be able to study and understand the worldview they normalize.
- E.g. the Aerial Violent Individual (AVI) Dataset.

Explainable Artificial Intelligence I

Black box AI

- Nowadays the majority of AI models are black box.
- We don't know why they give us the correct or wrong answer.
- This is one of the main reasons why AI is not widely used in sensitive fields (medicine).

Why an explanation is important

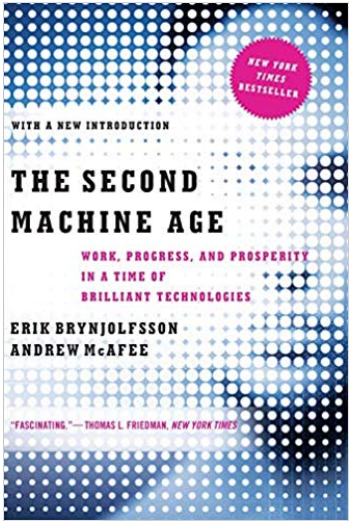
- We often wrongly explain the decisions of ML models based on our *human experience*. ML models might take decisions based on completely different reasons.
- E.g. a 2017 system tasked with image recognition learned to "cheat" by looking for a copyright tag that happened to be associated with horse pictures, rather than learning how to tell if a horse was actually pictured.

Explainable Artificial Intelligence II

Why an explanation is important

- Cooperation between human and AI should be based on trust.
- AI can find a shortcut in the training dataset.
- Accountability.
- Transparency.
- Corrigibility.

Lack of Jobs in the Age of AI I

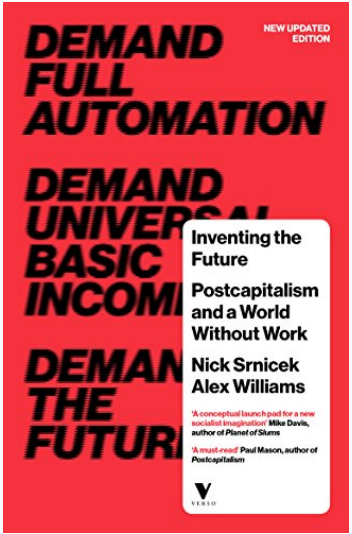


The AI revolution is different than any other technological revolution:

- AI can overcome human skills in a broad range of (high level) tasks (e.g. interpreting radiographs).
- Build a system that overcome human ability requires a small group of people and a short amount of time (Alphastar is developed by a team of 40 people in $\sim$1y)
- Once produced, the system can be replicated without any further cost.

Food for thought $\rightarrow$ what if the system is capable of learning continuously, so it doesn't need to be "updated"?

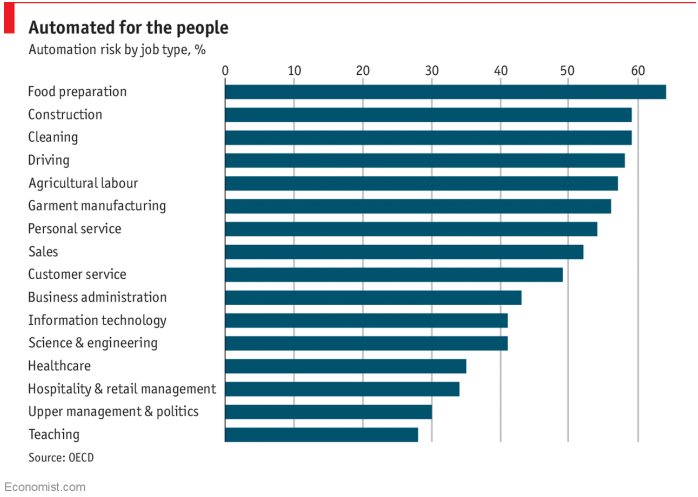
Lack of Jobs in the Age of AI II



The AI revolution is different than any other technological revolution:

- Not only repetitive or low level work will be replaced by machines.
  - Artificial creativity.
  - Computer-aided diagnosis.
  - Marketing.
  - Self-driving vehicles.
  - Developing AI?
- There is no place in the AI industry for all the replaced jobs!

Lack of Jobs in the Age of AI III



Adversarial Examples I

(a) Strawberry (b) Toy poodle (c) Buckeye (d) Toy poodle

Image from *"Generating Adversarial Examples with Adversarial Networks"*, *Xiao et al.*

Adversarial Examples II

Here's how scientists convinced self-driving cars that stop signs were speed limit signs

BY PATRICK CAIN - GLOBAL NEWS

Posted August 8, 2017 1:30 pm
Updated August 8, 2017 1:36 pm

— Subtle changes to traffic signs cause self-driving cars to "misbehave in unexpected and potentially dangerous ways," scientists have found. *arXiv*

Adversarial Examples III

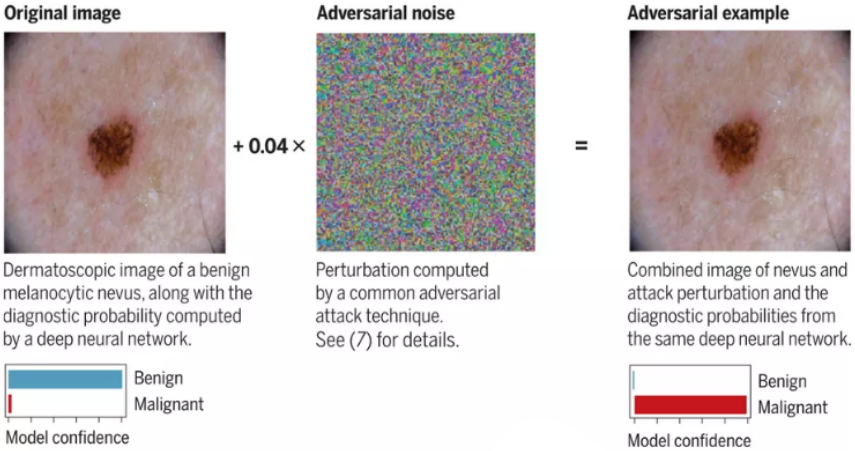


Image from *"Adversarial attacks on medical machine learning"*, *Finlayson et al.*

How to Build Ethical Machines

Risk Management & Safety I

- We saw that AI poses a wide range of issues and risk in both the near and the far future.
- Many people have advised us about the potential risks of AI (Bostrom, Kurzweil).
- Governments and big tech companies sponsor initiatives to find solutions to these problems.

So, what we are concretely doing to mitigate the risk and increase the safety of AI?

Risk Management & Safety II

Risk Management & Safety III

Why so empty?

- Industrial history:
  - Software development: not economically valuable
    - Just roll out v. 1.1 or a security patch → Actually perceived as valuable.
    - Cyber rather than physical.
  - Compare with infrastructure or industrial products engineering:
    - Product recalls and failures.
    - Very physical.
- Hype and startuppy culture:
  - Software development: minimise time to market, "easy money"/smart potato, clients' care.
  - Infrastructure and industrial products: entrenched industry, public tenders, educate the client.

Three layers of AI (and technology) Safety

- First Layer: Alignment
  - Do what I mean given this environment.
  - Technology works in intended use-cases.
  - E.g. bias and fairness.
- Second Layer: Robustness
  - Keep doing what I mean in unforeseen environment.
  - Technology is safe even in unintended use-cases.
  - E.g. ethics in decisions and adversarial attacks.
- Third Layer: Corrigibility
  - Enable me to detect and correct your mistakes.
  - Imperfect technology can be detected and improved over time.
  - E.g. white box models.

How to Insert Ethics in AI I

What is ethics?

- A review of 84 ethical AI documents by Jobin et al. (2019) found that no single principle featured in all of them.
- Nevertheless (!)
  - Themes of transparency, justice and fairness, non-maleficence, responsibility and privacy appeared in over half.
  - Themes of privacy, security, autonomy, justice, human dignity, control of technology and the balance of powers, were recurrent (Royakkers et al., 2018)

How to Insert Ethics in AI II

But...how?

- The guidelines often suggest that technical solutions exist, but very few provide technical explanations.
- 79% of tech workers report that they would like practical resources to help them with ethical considerations (Miller & Coldicott, 2019).
- Mapping needed to have a 'how' for every 'what'.

How to Insert Ethics in AI III

The most ethical approach?

- Ethics by design:
  - Can be paternalistic as it constrains the choices of agents.
  - i.e. speed bumps (permanent and leaves no real choice, especially in emergency).
- Pro-ethical design:
  - It does not preclude a course of action, but it requires the agents to make up their mind about it (still forces to make a choice, but less of a paternalistic nudge).
  - i.e. a speed camera (leave freedom to choose to pay a ticket, especially in emergency).

How to Insert Ethics in AI IV

The most ethical approach?

- Ethics by design:
  - Can be paternalistic as it constrains the choices of agents.
  - i.e. speed bumps (permanent and leaves no real choice, especially in emergency).
- Pro-ethical design:
  - It does not preclude a course of action, but it requires the agents to make up their mind about it (still forces to make a choice, but less of a paternalistic nudge).
  - i.e. a speed camera (leave freedom to choose to pay a ticket, especially in emergency).

Some Countermeasures I

Use explainable models

- An artificial intelligence model can be white box by design.
  - E.g. symbolic reasoning systems.
- We can theoretically know the output of the system for every possible input.
- We can inspect the system in order to find biases and weaknesses.
- A white box model is easier to fix.
- Explainability *a priori*.

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Some Countermeasures II

Explain black box models

- Attention models.
- Test the model with different data until the reasons of the input-output mapping is inferred.
  - E.g. cover portions of images until the most important patch is found.
  - E.g. change the data in a loan request until the bank's AI system accept/reject it.
- Explainability *a posteriori*.

Some Countermeasures III

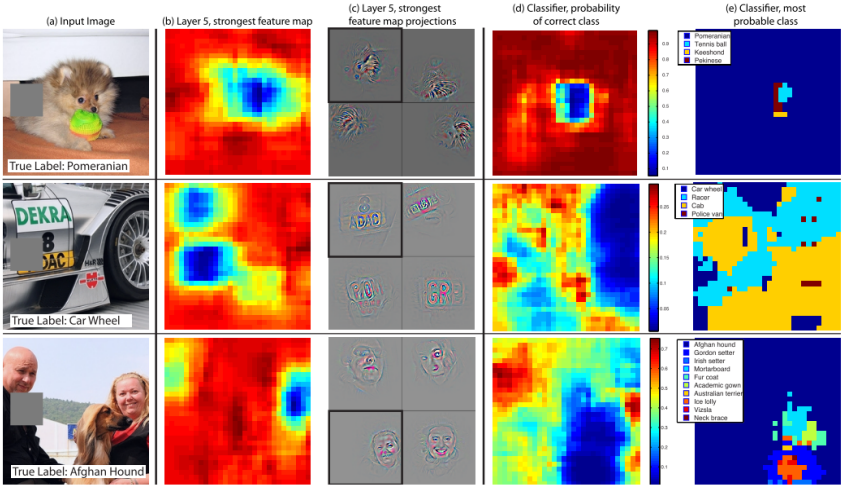


Image from *"Visualizing and Understanding Convolutional Networks"*, Zeiler et al.

Some Countermeasures IV

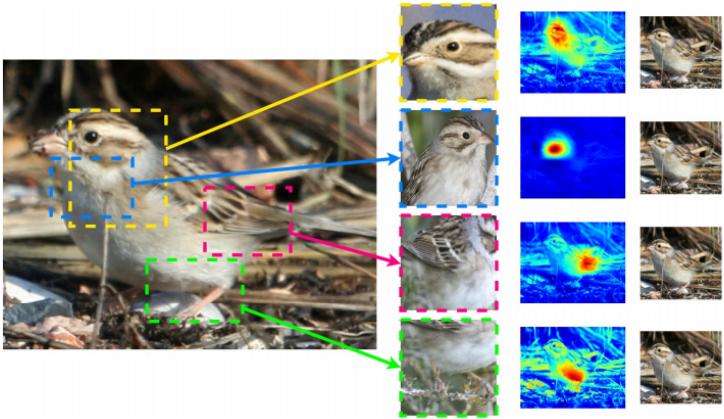


Image from *"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", Cynthia Rudin*

Some Countermeasures V

Democracy as countermeasure

Democratization of AI could be facilitated by providing equal access to:

- High-quality/high-quantity data.
- Cutting-hedge AI algorithms and research tools.
- Computational and engineering resources for ready-to-deploy solutions.

Homo Technologicus

Democratization of AI could be facilitated by providing equal access to:

- online identity which generates the need for and yet still lacks a new set of rights,
- People are the only owners of their data.
- Governments and policy regulations of AI and data.

AI for Social Good

Applying Artificial Intelligence for Social Good I

What is AI for social good?

- If one can address these technical issues, AI can fulfill many promises.
- What counts as AI for Social Good in practice?
- What makes AI socially good, in theory?

Must read → https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

Applying Artificial Intelligence for Social Good II

Image from *McKinsey Global Institute analysis*

AI for Social Good Applied to...

Crisis response

These are specific crisis-related challenges, such as responses to natural and human-made disasters in search and rescue missions, as well as the outbreak of disease. Examples include using AI on satellite data to map and predict the progression of wildfires and thereby optimize the response of firefighters. Drones with AI capabilities can also be used to find missing persons in wilderness areas.

Educational challenges

These include maximizing student achievement and improving teachers' productivity. For example, adaptive-learning technology could base recommended content to students on past success and engagement with the material.

AI for Social Good Applied to...

Environmental challenges

Sustaining biodiversity and combating the depletion of natural resources, pollution, and climate change are challenges in this domain. The Rainforest Connection, a Bay Area nonprofit, uses AI tools such as Google's TensorFlow in conservancy efforts across the world. Its platform can detect illegal logging in vulnerable forest areas by analyzing audio-sensor data.

Health and hunger

Researchers at the University of Heidelberg and Stanford University have created a disease-detection AI system—using the visual diagnosis of natural images, such as images of skin lesions to determine if they are cancerous—that outperformed professional dermatologists. AI-enabled wearable devices can already detect people with potential early signs of diabetes with 85 percent accuracy by analyzing heart-rate sensor data. These devices, if sufficiently affordable, could help more than 400 million people around the world afflicted by the disease.

AI for Social Good in Practice

- What counts as AI for Social Good in practice?
  - From solving a problem, making positive social impact...
  - ...To tackling its root causes!
- Tech Companies often lack the expertise in complex social and humanitarian issues
- As a solution...NGOs partnerships!
  - Intel partnered with National Geographic and the Leonardo DiCaprio Foundation on wildlife trafficking.
  - Facebook partnered with the Red Cross to find missing people after disasters.
  - IBM's social-good program alone boasts 19 partnerships with NGOs and government agencies.

AI for Social Good in Theory

- What makes AI socially good, in theory?
  - From good intentions...
  - ...To accountability!
- Analysing AI for Good ad hoc, by analysing specific areas of application, indicates the presence of a phenomenon
  - It does not explain how to ensure or reproduce its positive results.
  - AI for Good would benefit from an analysis of the essential factors that underline the design of a successful AI for Good system.

AI for People

AI for People

Our mission

Our mission is to learn, pose questions and take initiative on how Artificial Intelligent technology can be used for the social good. Our strategy is to conduct impact analysis, projects and democratic policies that act at the crossing of Artificial Intelligence and society. We are a diverse team of motivated individuals that is dedicated to bring AI Policy to the people, in order to create positive change in society with technology, through and for the public.

AI for People

Think tank

- Motivated people with heterogeneous background (cs, politics, ethics, philosophy, ai, ...)
- Monthly online meetup, when we discuss news, possible collaborations and ideas.
- Open to everyone!

Research

- *"Intelligent Drone Swarm for Search and Rescue Operations at Sea"*, presented at the "AI for Social Good" NeurIPS 2018 Workshop.
- We are currently working on a paper that defines and define strategies towards AI democratization.

AI for People

Education

- Blog and informal articles on Medium (https://medium.com/ai-for-people).
- Co-editor of Italian AI stories (https://medium.com/italian-ai-stories).
- We organize talks, seminars, hackathons, datathons etc.
  - Co-organizers Data Science Bologna Datathon.
  - Adviser for Deep Berlin AI for Good Hackathon.

Projects

- We are working on a SAR project, using drones and AI to detect boats.
- Computation, memory, transmission and many more problems.
- Join us if you are interested!

Contacts

- Website: https://www.aiforpeople.org
- Slack channel: click here!
- Mail: contact@aiforpeople.org

- Personal mail: gabriele.graffieti@gmail.com or gabriele.graffieti@unibo.it

Thank You!

Discussion?

Goal, Risks and Countermeasures in the Artificial Intelligence Era

DataScienceSeed #11

Gabriele Graffieti

PhD Student @ University of Bologna
Head of AI Research @ AI for People

January 23, 2020