

The Ethics of Artificial Intelligence



Prof. Roberto V. Zicari
Frankfurt Big Data Lab
www.bigdata.uni-frankfurt.de

March 6, 2019

ESOC, Darmstadt

Do no harm

Can we explain decisions?



What if the decision made using AI-driven algorithm harmed somebody, and you cannot explain how the decision was made?

- ❧ At present we do not really understand how Advanced AI-techniques such as used in Deep learning (e.g. neural networks) really works. It can be extremely difficult to understand which features of the data the machine used, and how they were weighted, to contribute to the outcome.
- ❧ This is due to the technical complexity of such advanced neural networks, which need huge amount of data to learn properly. It is a try and error.
 - ❧ This poses an ethical and societal problem.

Practically: When do we harm?



Accuracy

*“What happens if my algorithm is wrong?
Someone sees the wrong ad. What’s the harm?
It’s not a **false positive** for breast cancer.” (*)*

-- Claudia Perlich, Data Scientist

But Marketing/Social Media are not really harmless.
...and we do have fake news!
We also have **false negative**...

CR (*) Source: Big Data and The Great A.I. Awakening. Interview with Steve Lohr ODBMS Industry Watch, December 19, 2016

Harm: When things go wrong



Bias

When algorithms are used for example, to review loan applications, recruit new employees or assess potential customers, **if the data are skewed the decisions recommended by such algorithms may be discriminatory against certain categories or groups.**

Technically (*) Bias in machine learning= errors in estimation or over/under representing populations when sampling.

Selection, sampling, reporting bias

Bias of an estimator

Inductive bias

Other kinds of bias (**)

Allocative harm= when a system allocates or withholds a certain opportunity or resource

Representation harm = when a system reinforces the subordination of some groups along the lines of identity

(*) Source: CS 294: Fairness in Machine Learning UC Berkeley, Fall 2017 <https://mrtz.org/nips17/#/6>


(**) Source: Kate Crawford, Keynote "The Trouble with Bias" Neural Information Processing System Conference

Too homogeneous?



Diversity

“If AI/ML teams are too homogeneous, the likelihood of group-think and one-dimensional perspectives rises – thereby increasing the risk of leaving the whole AI/ML project vulnerable to **inherent biases and unwanted discrimination.**” -- Nicolai Pogadl (*)

 (*) Source: personal communication.

AI Safety



- ❧ “Deep neural networks can fail to generalize to **out-of-distribution inputs**, including natural, non-adversarial ones, which are common in real-time settings”. (*)
- ❧ “Several machine learning models, including neural networks, consistently misclassify **adversarial examples**---inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. (**)

(*) Source : **Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects** [Michael A. Alcorn](#), [Qi Li](#), [Zhitao Gong](#), [Chengfei Wang](#), [Long Mai](#), [Wei-Shinn Ku](#), [Anh Nguyen](#) (Submitted on 28 Nov 2018 (v1), last revised 13 Jan 2019 version, v2)

(**) Source: **Explaining and Harnessing Adversarial Examples** [Ian J. Goodfellow](#), [Jonathon Shlens](#), [Christian Szegedy](#) (Submitted on 20 Dec 2014 (v1), last revised 20 Mar 2015 version, v3)

AI and Democracy



"Big Nudging"

He who has large amounts of data can manipulate people in subtle ways. But even benevolent decision-makers may do more wrong than right.

☞ **Spotlight on China: Is this what the Future of Society looks like?**

How would *behavioural* and *social control* impact our lives? The concept of a Citizen Score, which is now being implemented in China, gives an idea.

Source: Will Democracy Survive Big Data and Artificial Intelligence?. Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A.. (2017). Scientific American (February 25, 2017).

Who is responsible?



AI system designers and their managers do have ethical responsibilities.

and

Other stakeholders (e.g. policy makers, politicians, opinion leaders, educators) do have ethical responsibilities.

Example : Autonomous Cars



Let`s consider an autonomous car that relies entirely on an algorithm that had taught itself to drive by watching a human do it.

What if one day the car crashed into a tree, or even worse killed a pedestrian?

The Uber Case for *False positive* for plastic bags...



„The newsletter "The Information" has reported a leak from Uber about their fatal accident. The relevant quote:

The car's sensors detected the pedestrian, who was crossing the street with a bicycle, but Uber's software decided it didn't need to react right away. **That's a result of how the software was tuned.** Like other autonomous vehicle systems, Uber's software has the **ability to ignore "false positives,"** or objects in its path that wouldn't actually be a problem for the vehicle, such as a plastic bag floating over a road. In this case, Uber executives believe the company's **system was tuned so that it reacted less to such objects.** But the tuning went too far, and the car didn't react fast enough, one of these people said." (*)

(*) How reliable is this Source? : <https://ideas.4brad.com/uber-reported-have-made-error-tuning-perception-system>

Story also in Der Spiegel Nr. 50/8.12.2018 *Tod durch Algorithms* (Philipp Oehmke)

Algorithms learn from data



“Since the algorithms learn from data, it’s not as easy to understand what they do as it would be if they were programmed by us, like traditional algorithms. But that’s the essence of machine learning: that it can go beyond our knowledge to discover new things. A phenomenon may be more complex than a human can understand, but not more complex than a computer can understand.

*And in many cases we also don’t know what humans do: for example, we know how to drive a car, but we don’t know how to program a car to drive itself. **But with machine learning the car can learn to drive by watching video of humans drive.**” (*)*

--- Pedro Domingos

(*) Source: *On Artificial Intelligence, Machine Learning, and Deep Learning*. Interview with Pedro Domingos, ODBMS Industry Watch, June 18, 2018

Learning the Good and the Bad



Waymo (a subsidiary of [Alphabet Inc](#)) created a Recurrent Neural Network (RNN) for Driving.

They trained the neural network Imitating the “**Good**” and synthesizing the “**Bad**”.

„They trained the model with examples from the equivalent of about 60 days of **expert driving** data, while including training techniques such as past motion dropout to ensure that the network doesn't simply continue to extrapolate from its past motion and actually responds correctly to the environment.“

(*) Source : **Learning to Drive: Beyond Pure Imitation**
Dec 10, 2018, <https://medium.com/waymo/learning-to-drive-beyond-pure-imitation-465499f8bcb2>

ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst. Mayank Bansal, Alex Krizhevsky, Abhijit Ogale, Dec 7, 2018 <https://arxiv.org/pdf/1812.03079.pdf>

Learning from “Bad Examples”



“It's not difficult to feed the bad examples. That's what we do in our training, we feed it synthesized bad examples and add a training loss that tells the network not to emulate the bad behavior.

Real examples of bad behavior are difficult to intentionally obtain, and it is **simpler** and **safer** to **synthetically** create bad examples in **simulation**.”

--[Abhijit Ogale](#)

Learning: Who sets the examples?



“If the learning took place before the car was delivered to the customer, the car’s manufacturer would be liable, just as with any other machinery. The more interesting problem is if the car learned from its driver.

Did the driver set a bad example, or did the car not learn properly?”

--Pedro Domingos

(*) Source: *On Artificial Intelligence, Machine Learning, and Deep Learning*. Interview with Pedro Domingos, ODBMS Industry Watch, June 18, 2018

Machine Learning and Causality



- ❧ **Causality** – in other words, grasping not just patterns in data but why something happens. Why is that important, and why is it so hard?
- ❧ “ If you have a good causal model of the world you are dealing with, you can generalize even in unfamiliar situations. That’s crucial. We humans are able to project ourselves into situations that are very different from our day-to-day experience. **Machines are not, because they don’t have these causal models.**
- ❧ We can hand-craft them, but that’s not enough. **We need machines that can discover causal models.** To some extent it’s never going to be perfect. We don’t have a perfect causal model of the reality; that’s why we make a lot of mistakes. But we are much better off at doing this than other animals.
- ❧ **Right now, we don’t really have good algorithms for this**, but I think if enough people work at it and consider it important, we will make advances.”

-Yoshua Bengio

- ❧ (*) Source MIT Technology Review
<https://www.technologyreview.com/s/612434/one-of-the-fathers-of-ai-is-worried-about-its-future/>

The WHY question



”Knowing **why** an expert driver behaved the way they did and what they were reacting to is critical to building a causal model of driving. For this reason, simply having a large number of expert demonstrations to imitate is not enough. **Understanding the why** makes it easier to know how to improve such a system, which is particularly important for safety-critical applications.” (*)

However, I do not believe that we know WHY and HOW we drive though...

Try for yourselves: Explain to another person how do you drive and why you react in certain situations they way you do..... And please let me know the result.

(*) Source : **Learning to Drive: Beyond Pure Imitation**
<https://medium.com/waymo/learning-to-drive-beyond-pure-imitation-465499f8bcb2>

Intelligence and Ethical behavior



- ❧ “As a layperson looking at this particular field of ethical systems, I see some parallels between determining whether a system has intelligence and whether a system is making ethical decisions or not. In both cases, we are faced with a **kind of Turing test scenario** where we find it difficult to articulate what we mean by intelligence or ethics, and can only probe a system in a Turing test manner to determine that it is indistinguishable from a model human being.
- ❧ The trouble with this approach though is that we are assuming that if the system passes the test, it shares the same or similar internal representations as the human tester, and it is likely that its intelligence or ethical behavior generalizes well to new situations. We do the same to assess whether another human is ethical or not.
- ❧ **This is a great difficulty, because we currently know that our artificial ML systems learn and generalize differently than humans do, so this kind of approach is unlikely to guarantee generally intelligent or ethical behavior.**
- ❧ **I think the best we can currently do is to explicitly engineer/bound and rigorously test the system against a battery of diverse scenarios to check its decisions and reduce the likelihood of undesirable behavior.**
- ❧ **The number of tests needs to be large and include long-tail scenarios because deep learning systems don't have as large a generalization horizon as human learning, as evidenced by their need of a mountain of training data. “**

--- [Abhijit Ogale](#)

Disclaimer: personal viewpoint as a ML researcher, not in his role at Waymo.



*“Citizens and businesses alike need to be able to **trust** the technology they interact with, and have effective safeguards protecting fundamental rights and freedoms.*

*In order to increase **transparency** and **minimise the risk of bias**, AI systems should be developed and deployed in a manner that allows humans to **understand** the basis of their actions.*

*“**Explainable AI** is an essential factor in the process of strengthening people’s trust in such systems.” (*)*

-- Roberto Viola

Director General of DG CONNECT (Directorate General of Communication Networks, Content and Technology) at the European Commission.

☞ (*) Source [On the Future of AI in Europe. Interview with Roberto Viola](#), ODBMS Industry Watch2018-10-09

Accountable AI?



- ❧ Do we need some sort of auditing tool?
- ❧ The technology has to be able to “**explain**” itself, to explain how a data-driven algorithm came to the decision or recommendation that it did. Is it technically feasible?

This is current research work area: e.g

Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi (Submitted on 14 Jan 2019)

- ❧ How much **Transparency** is desired/ possible/ allowed....?
- ❧ Do we wish “**Human in the loop**” for most of these kinds of decisions for the foreseeable future?

AI and The Paradox of Transparency



- ❧ I do not mean *cognitive biases*...
- ❧ I mean, if we really insist on *AI Transparency*, perhaps this would force us to reveal our real *motives*...
- ❧ But, we do not always wish to make our motives visible to the outside world, e.g. we do not wish transparency....
- ❧ But with no transparency, there is a lack of trust.

(Non) Ethical People and (Non) Ethical AI



*“I think ethical software development for AI is not fundamentally different from ethical software development in general. The interesting new question is: **when AIs learn by themselves, how do we keep them from going astray?**”*

*Fixed rules of ethics, like Asimov’s three laws of robotics, are too rigid and fail easily. (That’s what his robot stories were about.) **But if we just let machines learn ethics by observing and emulating us, they will learn to do lots of unethical things.***

So maybe AI will force us to confront what we really mean by ethics before we can decide how we want AIs to be ethical.” ()*

--Pedro Domingos (*Professor at University of Washington*)



(*) Source: **On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos**, ODBMS Industry Watch, June 18, 2018

Testing and validating AIs



Perhaps we can “certify” AIs by the number of testing with synthetics data sets and extreme scenario they went through before allowing AIs to drive a car (similar to what happens to airplane pilots).

Somebody would need to define when good is enough. And this may be tricky...

☞ More feedback I have received, and resources here:

www.odbms.org/blog/2018/10/big-data-and-ai-ethical-and-societal-implications/#comments

AI Ethics inside?



Can Ethics be "embedded" into the core of the AI design?

Not reacting to it....

Kind of "*Ethics inside*".

The Need of AI Ethical Due Diligence?



They have used 30 million of real-world “expert” driving examples.

Q. How did you define an “expert” in your work? In the blog they speak of a “good driver”. Who is a good driver?

In the paper they write that “this is a difficult robotics challenge (i.e. predict, plan) that humans solve well”.

Q. How do you define if a human solve this well? This is particularly true when logically unexpected and unpredictable things happens on the road (a fire, a hearth quake, a bridge collapses, etc.)

They design a RNN to output a trajectory which consists of ten future points.

Q. Why ten? Any particular rational for this?

They define an “imitation dropout” as composed of imitation losses plus environment losses.

Q. How is the learning (and accuracy) affected if you change the dropout strategy?

What I am interested in..



People motivation plays a key role here. With AI the important question is how to avoid that it goes out of control, and how to understand how decisions are made.

What I am interested in (no particular order) :

- ☞ Raise awareness;
- ☞ Help create and support best practices- bear with me for the moment, for lack of a better term- I call them *"holistic AI Ethics eco-systems"*;
- ☞ Talk to AI developers, learn if Ethics can be "embedded" into the core of the AI design, kind of *"Ethics inside"*;
- ☞ Learn what measures have to be taken for achieving a trustful AI;
- ☞ Write a book on AI and Ethics for the general public.

Back to the Future...



Are we talking about Science Fiction here or...

“The ideal of **General AI** is that the system would possess the cognitive abilities and general experiential understanding of its environments that we humans possess, coupled with the ability to process this data at much greater speeds than mortals. It follows that the system would then become exponentially greater than humans in the areas of knowledge, cognitive ability and processing speed – giving rise to a very interesting species-defining moment in which the human species are surpassed by this (now very) strong AI entity.”

Source: <https://hackernoon.com/general-vs-narrow-ai-3d0d02ef3e28>

and this poses severe Ethical concerns....

I also believe that AI initiative such as **Neuralink**: <https://www.neuralink.com/> poses serious Ethical issues...

*„Creating a neural lace is the thing that really matters for humanity to achieve symbiosis with machines” --**Elon Musk** (*)*

<https://www.cnbc.com/2018/09/07/elon-musk-discusses-neurolink-on-joe-rogan-podcast.html>

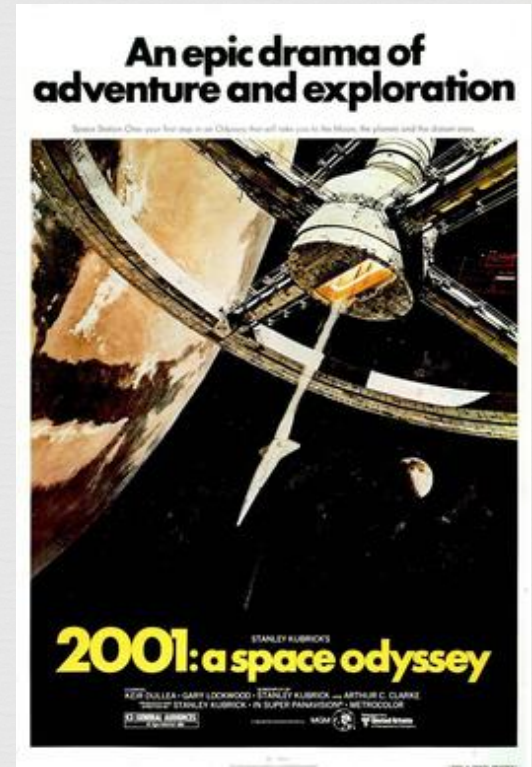
AI in Space?




2001: A Space Odyssey

is a 1968 science fiction novel by British writer Arthur C. Clarke. It was developed concurrently with Stanley Kubrick's film version and published after the release of the film.

Source: Wikipedia.



Tower of Babel



And they said, Go to, let us
build us a city and a tower,
whose top may reach unto
heaven;

— *Genesis 11:4*

The Tower of Babel by
Pieter Bruegel the Elder(1563)