Machine Learning and Statistical Learning

Machine Learning and Ethics

Ewen Gallic
ewen.gallic@gmail.com

MASTER in Economics - Track EBDS - 2nd Year

1. Machine Learnign in Everyday Life: a (not Complete) Overview

To conclude this course

As we saw in the introduction of this course, **machine learning** techniques are used to learn a **mapping** from the **data** to a **prediction**, without explicitly programming the learning rules.

We have mainly talked about **supervised learning** in this course, and focused on two kinds of tasks:

- regression
- classification

We mentioned that the aim of those tasks could be:

- to make a correct prediction
- to explain some prediction

In this last part of the course, let us focus on an aspect that was left aside for now: **ethics**.

Machine Learning is Everywhere

Before presenting the ethical questions raised by machine learning, let us have a quick overview of different fields in which machine learning is currently used and has expanded during the last decades.

1.1 Agriculture

Agriculture

Agriculture represents a non negligible part of the economy, even in relatively developped countries.

Profits in the agricultural sector depend on agricultural yields, which in turn depend on many factors, such as the fertility of the land or weather conditions.

The use of machine learning algorithms in the agricultural sector is growing rapidly. According to Daniel Faggella (head of Research at Emerj), artificial intelligence in agriculture can be categorized into three categories (see AI in Agriculture – Present Applications and Impact)

1. Agricultural robots
2. Crop and soil monitoring
3. Predictive analytics

1.1.1 Agricultural robots

Agricultural robots

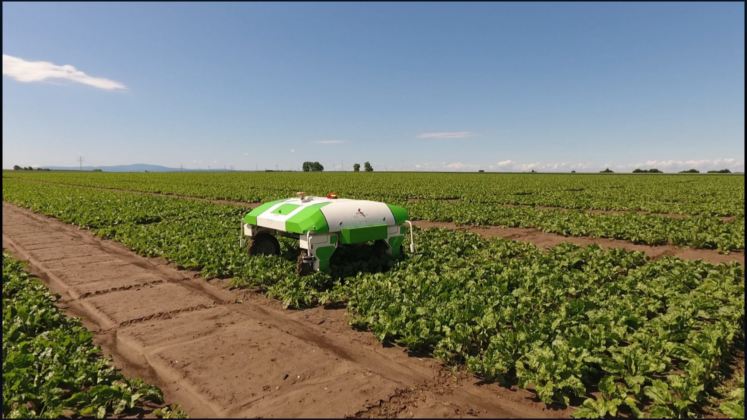


Figure 1: FieldBot in sugar beet field. Source: Europeanseed

Industrial robots used to

- spray pesticides / herbicides
- harvest crops

1.1.2 Crop and soil health monitoring

Crop and soil health monitoring

The agricultural industry faces serious threats related to (among others):

- soil erosion
- pests

These affect crop production and may turn to dramatic food security issues. To monitor crop and soil health, some algorithms are used to identify problems related to:

- soil conditions (dryness, lack of nutrients)
- presence of plant pests, plant diseases

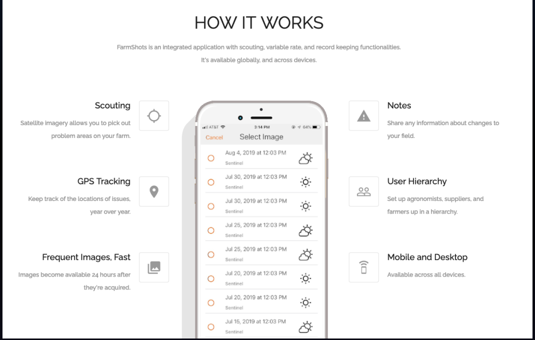
The agricultural sector relies more and more on drones or satellites images (not only for crops, but also for cattle monitoring).

1.1.3 Predictive analytics

Predictive analytics

Crop yields closely depend on weather conditions. Forecasting accurately the weather is therefore of prime importance for farmers.

- the success of a harvest can be strongly affected by the times at which planting or harvesting takes place
- these dates may in turn be affected by weather conditions.



HOW IT WORKS

FarmShots is an integrated application with scouting, variable rate, and record keeping functionalities.
It's available globally, and across devices.

Scouting
Satellite imagery allows you to pick out problem areas on your farm.

GPS Tracking
Keep track of the locations of issues, year over year.

Frequent Images, Fast
Images become available 24 hours after they're acquired.

Notes
Share any information about changes to your field.

User Hierarchy
Set up agronomists, suppliers, and farmers up in a hierarchy.

Mobile and Desktop
Available across all devices.

Figure 2: Farmshots uses satellites data to monitor crop health. Source: Farmshots.com

1.2 Education

Education

Machine learning can be helpful in the Education field (see The Role Of Artificial Intelligence In The Classroom):

- Helping grading the students
- Providing individualised teaching (more details in the next slide)
- Predicting career paths
- Improving accessibility

Education

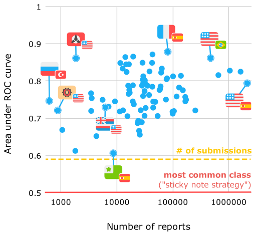
According to Acemoglu and Restrepo (2019), AI:

- has a very low penetration in Education
- may be used to provide students with **individualised teaching**
- could provide students with skills that will be needed in the future (rather than the backward-looking curricula)
- would even create jobs (not only to develop the algorithms, but to put the individualised teaching as well)

Education: the case of Duolingo

The American language-learning website “Duolingo” uses machine learning algorithms to improve their service and to prioritize course improvements.

- How machine learning helps Duolingo prioritize course improvements
- personalisation of learning (See Settles and Meeder (2016) and Medium article by Fayrix Software)



Prioritizing reports for all of our courses: Each language direction (”<big flag> for <little flag> speakers”) plotted by how much data we have and how accurate the machine learning system is for it. Source: Duolingo.

1.3 Military

Military

Machine Learning is also widely used in the military sector.

According to Wong et al. (2020), three trends explain why AI tends to be more and more used in the military:

1. Since 11 September 2001, the number of unmanned systems (mostly controlled by humans) rose on the battlefield (aerial systems, including drones, ground robots)
2. Huge advances in research in AI
3. The will of major powers to use AI autonomous systems in their millitaries

Some of the applications concern:

- assistance in decision making (quick process using data collected)
- detecting hidden targets

1.4 Finance

Finance

The financial world also relies on many machine learning algorithms. Some of the uses are the following:

- Management of personal Finance (prodiving individualised assistance)
- Fraud prevention
- Prediction of loan risks (risk assessment)
- Predicting the future value of assets

1.5 Health

Health

The healthcare field is benefiting from the growing research in machine learning coupled with access to massive volumes of data:

Davenport and Kalakota (2019) provides an overview of what ML is used for in healthcare:

- Precision medicine, for predicting the success of treatment protocols for patients
- NLP to analyse clinical notes, to prepare reports, to transcribe patient interactions
- Rule-based expert systems
- Physical robots
- Robotic process automation to complete administrative tasks
  [Machine Learning] is a natural extension to traditional statistical approaches. Beam and Kohane (2018)

Health: monitoring diseases

Machine learning techniques are also widely used to monitor diseases, such as the Malaria, Zika, or Ebola:

- See Rogers et al. (2002) for study and forecast of Malaria using satellite images

1.6 Human resources

Human resources

- Matching employers and job seekers on the job market
  - scouting for candidates
  - automatically screening resumes
  - predicting attrition

2. Research Questions about Ethics in Machine Learning

Research Questions about Ethics in Machine Learning

Among these applications of machine learning, some of them, although widely used to try achieve their objectives, also raise **ethical issues**.

A recent paper (Piano 2020)) provides a really nice overview of ethical principles in machine learning and artificial intelligence.

Let us take some time to consider some issues relative to machine learning, then let us look at some definitions of morals and ethics, and then, let us have a look at the current research questions about ethics in machine learning.

Question

But first, I would like to get your opinion.

Potential broad issues of Machine Learning

The following questions seem to emerge (see Piano (2020)):

- Machine learning models rely on **correlations** to provide predictions:
  - potential overfitting problems, detection of **black swans**
  - and more importantly here, variables used may be **proxies** for driving trends that may lead to potential **discrimination issues**
- What degree of autonomy should be left to the algorithms?
- How can we avoid ethical issues?
  - *ex-ante* evaluation (incorporating some rules, *e.g.*, by law, when creating the algoritm)
  - *ex-post* evaluation (monitoring the consequences and then adjusting)

Potential broad issued of Machine Learning

More on **black swans**:

> *a black swan is a highly improbable event with three principal characteristics: It is unpredictable; it carries a massive impact; and, after the fact, we concoct an explanation that makes it appear less random, and more predictable, than it was. Taleb (2007)*

Three attributes:

- rarity
- extreme impact
- retrospective predictability

Ethics: definition

According to Cointe, Bonnet, and Boissier (2016):

> ***Morals*** *consists in a set of moral rules which describes the compliance of a given behavior with mores, values and usages of a group or a single person. These rules associate a good or bad value to some combinations of actions and contexts. They could be specific or universal, i.e. related or not to a period, a place, a folk, a community, etc.*
>
> ***Ethics*** *is a normative practical philosophical discipline of how humans should act and be toward the others. Ethics uses ethical principles to conciliate morals, desires and capacities of the agent.*

Ethics: three approaches

According to Cointe, Bonnet, and Boissier (2016), three major approaches are considered in the literature:

- **Virtue ethics**
  - an agent is ethical if and only if he or she **acts and thinks according to some moral values** (wisdom, bravery, justice, ...)
- **Deontological ethics**
  - an agent is ethical if and only if he or she **respects obligations and permissions** related to possible situations
- **Consequentialist ethics**
  - an agent is ethical if and only if he or she **weights the morality** of the consequences of each choice and chooses the option which has the most moral consequences

Ethics: ethical dilemnas

As pointed out by Yu et al. (2018), **ethical dilemmas** occur when it is necessary to break an ethical principle in order to make a decision that must necessarily be made

- for example, in case of a car accident involving a driverless car, as explained in Kirkpatrick (2015).

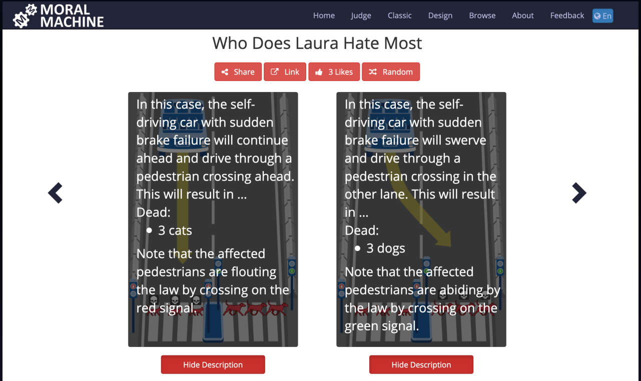
Ethics: ethical dilemmas

Taxonomy proposed by Yu et al. (2018) to divide the field of AI and ethics into four areas:

1. **exploring ethical dilemmas**: to understand **human preferences** on ethical dilemmas
2. **individual ethical decision frameworks**: decision-making mechanisms in which an **individual agent** can judge the ethics of his or her own actions and the actions of other agents
3. **collective ethical decision frameworks**: decision-making mechanisms in which **multiple agents** can take a collective decision
4. **ethics in Human-AI interactions**: including ethical considerations into agents which are designed to **influence human behaviours**.

Ethics: exploring ethical dilemmas

To take a collective decision, some frameworks have been implemented online, such as the Moral Machine from the MIT



Who Does Laura Hate Most

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...
Dead:

- 3 cats

Note that the affected pedestrians are flouting the law by crossing on the red signal.

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...
Dead:

- 3 dogs

Note that the affected pedestrians are abiding by the law by crossing on the green signal.

Figure 3: Example of a scene to judge on the Moral Machine

Ethics: exploring ethical dilemmas

The self-reported preferences of 3M participants in the Moral Machine Project highlighted that: (Yu et al. 2018)

- People have libertarian ideas: autonomous vehicles should make sacrifices if it saves more lives.
- If an autonomous vehicle can save more pedestrian lives by killing its passenger, people tend to prefer *others*' vehicle to have this feature rather than their own

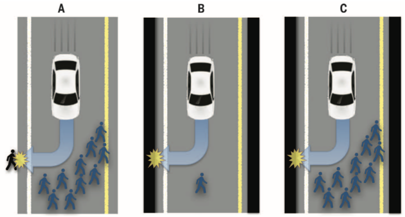


Fig. 1. Three traffic situations involving imminent unavoidable harm. The car must decide between (**A**) killing several pedestrians or one passerby, (**B**) killing one pedestrian or its own passenger, and (**C**) killing several pedestrians or its own passenger.

Figure 4: Three traffic situations involving imminent unavoidable harm.

Source: Bonnefon, Shariff, and Rahwan (2016)

Ethics: exploring ethical dilemmas

- Study 1 : sacrifice one passenger vs kill 10 pedestrians
- Which would be the most moral way to program AVs?
  - preference for AVs programmed to kill their passengers for the greater good

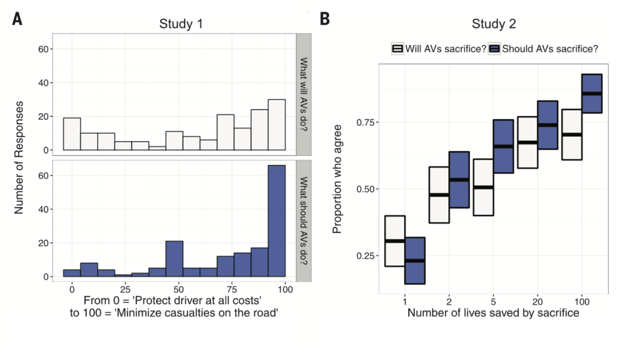


Figure 5: Considering the greater good versus the life of the passenger.

Source: Bonnefon, Shariff, and Rahwan (2016)

Ethics: ethics in Human-AI interactions

Some Artificial Intelligence may attempt to modify our beahaviour. According to Yu et al. (2018), these modifications should respect the three principles of The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979):

- **Respect for persons**: individuals should be treated as **autonomous agents**
- **Beneficience**: beneficent actions should **not harm**, should **maximize** the possible **benefits** and **minimize** possible **harms**
- **Justice**: **fairness distribution** of the benefits and risks among the users

3. What Should we Care About?

What should we care about?

Jobin, Ienca, and Vayena (2019) analysed 84 (recent) documents containing ethical principles or guidelines for AI. They summarised the values and principles according to the following values:

- **transparency**
- **justice, fairness and equity**,
- **non-maleficence**
- **responsibility and accountability**
- **privacy**
- freedom and autonomy
- trust
- dignity
- sustainability
- solidarity

The five highlighted principles appeared in more than half of the documents.

3.1 Transparency

The GDPR and transparency

In Europe, transparency has become an important principle in Machine Learning under the General Data Protection Regulation (GDPR), *i.e.*, the data protection law that passe din 2016 and came into operation in Europe in 2018.

The Article 5.1.a of the GDPR, titled "Principles relating to processing of personal data", states:

> *Personal data shall be: processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')*

According to Felzmann et al. (2019), we need to consider transparency in two manners here:

- **prospective transparency**: individuals must be informed **prior** the estimation (see also article 12)
- **retrospective transparency**: the decisions made/reached must be **traceable**

The GDPR and transparency

The Article 22 of the GDPR, titled "Automated individual decision-making, including profiling" states:

> *The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*

The Article 15 1 h, titled "Right of access by the data subject":

> *the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.*

The GDPR and transparency: issues

The article 22 (and the article 15) could imply that individuals could ask for a **right to an explanation**.

However, as noted by Edwards and Veale (2018), the article 22 has **several problems**:

1. article 22 "*is merely a right to stop processing unless a human is introduced to review the decision on challenge*' '.
2. it only concerns **significant** decisions made by algorithms where there is **no human involved**
3. the term **decision** is not coined explicitly

But in the end, what is transparency?

The key elements reported in Jobin, Ienca, and Vayena (2019) about transparency are the following:

- Explainability
- Interpretability
- Communication

These concern:

- the **algorithms** used
- the **predicted values**

Regarding the interpretability, Rudin (2019) argues that one should develop **interpretable models** in the first place instead of “black box” models when facing high stakes decisions.

Could Open Source Softwares be a Solution to More Transparency?

A solution to an increased transparency could be to resort to **Open Source softwares**.

According to Sonnenburg et al. (2007):

1. **Reproducibility** of scientific research is increased
2. Algorithms implemented in same framework facilitate **fair comparisons**
3. **Problems can be uncovered** much faster
4. Bug fixes and extensions from external sources
5. Methods are **more quickly adopted** by others
6. **Efficient algorithms** become available
7. Leverage existing resources to aid new research
8. Wider use leads to wider recognition
9. More complex machine learning algorithms can be developed
10. Accelerates research
11. Benefits newcomers and smaller research groups

Could Open Source Softwares be a Solution to More Transparency?

But on the other hand:

1. **High entry costs** :
   - existing methods are already **complex**
   - algorithms already restricted to a few group of people, newcomers have to redo the work of others first
2. Machine Learning Researchers are **not good programmers**
3. **Low incentive** for publishing in Open Source

Could Open Source Softwares be a Solution to More Transparency?

Some other reasons are put forward by Laat (2017):

1. Loss of **privacy** when private data are exposed
   - as explained by Zarsky (2013), may lead to stigmatization when people cannot escape from some of their characteristics such as their race, religion, ethnicity, nationality and gender
2. Possibility to "game" the system (*e.g.*, online reviews), to manipulate it (**perverse effect**)
3. Opacity may be needed for **competitive** reasons between firms (**intellectual property** issues)
4. Opacity, interpretability and explainability: as **algorithms are usually complex**, more transparency does not really help in that direction

3.2 Justice, Fairness and Equity

Bias and Fairness

Let us have a look at three examples in this part:

1. Bank loan qualification
2. Recidivism predictive algorithms
3. A simple scene detection algorithm

Through these examples, we will talk about justice, fairness and equity.

3.2.1 Bank Loan Qualification

Setup

Let us consider the **credit market**, in which agents with the **capacity to finance** lend money to other agents in **need of financing**.

The lender hopes to be able to be **repaid** one day and **avoid any credit default**

He or she faces a problem of **information asymmetry**: he or she has *a priori* less information than the borrower on the borrower's capacity to honour his or her debt.

Solution: create a model that can assess the risk of default on a loan.

The model

What are the variables that are included into that model? (some of these variables are not allowed to be used in Europe)

1. variables that describe the **borowwer's finance** (level of indebtedness, delays in reimbursements, payment of bills on time, ...)
2. variables that describe the **personal characteristics** or those of the **environment** (age, gender, race, zip code, religion, ...)

Among the variables in the second group: most of those are **proxies** and actually describe the **group** of persons that share similar characteristics.

The model

As explained in Lee and Floridi (2020), the model may **over-estimate risk** or **under-estimate** it.

The authors identidy three possible sources of over-estimation of minority risk:

1. **Selection bias**
   - may result from the difference in the advertisement of loans
   - low targetted ads in high-minority areas
   - low application from these areas
   - over-representation of some individuals with possibly specific characteristics from these ares
2. **Disparate treatment** (direct discrimination)
   - based on group belonging (statistical discrimination: group statistics used to judge an individual)
   - or on preferences

Over-estimation

3. **Self-perpretuation of the selection bias**
    - “nasty feedback loop” (O’neil 2016): as there is no counterfactual avaible on the candidates that were excluded (the loan was rejected), the model cannot learn from those individuals

Consequences :

- incorrect training dataset
- over-estimation of the risks for minorities
- under-estimation of the risks for non-minorities

What could be done?

We face here a **trade-off between fairness** (demanded by society) and **algorithmic accuracy** (demanded by private sector).

People from minorities are discriminated against when applying for credit. This creates both problems of selection bias and statistical discrimination, leading to **an over-estimation of the risk of default**.

As a solution, Lee and Floridi (2020) suggest the following:

- expand advertising activities aimed at ethnic minorities to limit self-selection issues
- lend more to minority people to be able to observe counterfactuals in the future and thus improve the predictive capabilities of the algorithm.

While in the **short-run**, this may reduce the predictive capacities of the model, it is not the case in the long-run. In addition, it may increase the market share.

3.2.2 Biased algorithms in Justice

KW 120578 AMG
04.15.2054
14:34:56
SS/MR/02
A STEVEN SPIELBERG FILM
MINORITY REPORT
TOM CRUISE COLIN FARRELL SAMANTHA MORTON AND MAX VON SYDOW JOHN WILLIAMS
INDUSTRIAL LIGHT & MAGIC DEBORAH L. SCOTT MICHAEL KAHN, A.C.E.
ALEX McDOWELL JANUSZ KAMINSKI, ASC GARY GOLDMAN RONAL
GERALD R. MOLEN BONNIE CURTIS WALTER F. PARKES JAN DE BONT
PHILIP K. DICK SCOTT FRANK AND JON COHEN STEVEN SP

Figure 6: Minority Report, Steven Spielberg (2002).

Situation

Let us have a look at a case study about an algorithm that **attributes a score** to defendants to estimate the **likelihood of them committing a crime again** if they are released from prison.

This kind of algorithm is used in the USA at different stages of the criminal justice system.

It is based on the predicted score, **decisions are made by humans**.

The idea is to **guide humans** into the sentence that can be given to defendants:

- it can help decide whether or not a prisoner should get an early release from prison
- it should help to limit the number of errors made by judges by providing an **objective measure of risk**

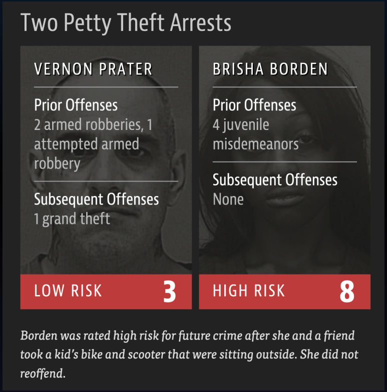
Situation

The objectives are thus two-fold:

1. Assessing the risk
2. Reducing the risk

What happens when the algorithm is built on biases from the data? Let us have a look.

Case study



Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Figure 7: Likelihood of two people to commit a future crime.

Source: Angwin et al. (2016)

- The algorithm (COMPAS) attributes a score to each convicted individual, to estimate the likelihood of them committing a crime again if they are released from prison.
- The individual with a low score committed a new crime
- The individual with a high score did not
- The output of the algorithm pointed to a wrong direction. What happened?

What are the results obtained with this algorithm?

- Angwin et al. (2016) (ProPublica) looked at the algorithm used
- The dataset they look at contains $n = 7,000$ cases of people arrested in Broward County, Florida, between 2013 and 2014
- They can compare the score assigned by the algorithm and whether or not these people committed a crime within the next two years
- Results:
  - only 20% of people predicted to commit violent crimes did so
  - only 61% of people predicted to commit a crime (regardless of its severity) actually did so

What are the results obtained with this algorithm?

What about race?

- the rate of accuracy (overall accuracy) is the same for Black and for White people: 63.8% (Dressel and Farid 2018)
- but if one looks at the false positive and false negative depending on race: **the error rate of the algorithm bas completely biased**
  - **False positive** (incorrectly predicting defendants who did not recidivate)
    - Black people: 44.9%
    - White people: 23.5%
  - **False negative** (incorrectly predicting defendants who did recidivate)
    - Black people: 28.0%
    - White people: 47.7%
- note: the owning firm of the algorithm (Northpointe) disputed the analysis, which was later on disputed by ProPublica

What variables does the algorithm use?

How does the algorithm work?

- The score is made of 137 variables (either questions answered by the defendants or extracted from their criminal record)
  - "Was one of your parents ever sent to jail or prison?"
  - "How many of your friends/acquaintances are taking drugs illegally?"
  - "How often did you get in fights while at school?"
  - Agree/disagree questions:
    - "A hungry person has a right to steal"
    - "If people make me angry or lose my temper, I can be dangerous.
- The variable about "race" is not included in those

What variables does the algorithm use?

- There are, however, numerous **proxies** for it, which may create the biases:
  - measure of poverty (poor neighborhoods)
  - measure of unemployment (neighborhoods with high unemployment rate)
  - measure of education (neighborhoods with terrible schools)

How come these bias may appear?

- As pointed out by Park (2019), the area of residence may be a proxy for race
  - if the police is targetting relatively more areas in which persons of colour live (as explained in O'neil (2016), chapter 5), the rate of recidivism will be relatively higher in these areas
  - as a consequence, Black people would then be relatively more often associated with high risk of recidivism.
- The training dataset would in that case associate high values of recidivism to areas where the proportion of Black individuals is relatively high

So, how to avoid this?

- The predictions of such a model should be **interpretable**
- **Publishing** the codes of the algorithms and the data allows for independent testing, so that biases may be discovered
  - publishing such sensitive data about individuals is subject to controversy
- If the data acquired creates “nasty feedback loop” (new data that come as a consequence of the prediction of the algorithm and reinforce the bias) and enhance inequalities: O’neil (2016) suggests to simply **drop that data**
- Do we really need this kind of model?
  - Dressel and Farid (2018) showed that similar results (in terms of fairness – or lack of– and accuracy) can be obtained when asking a crowd to make the prediction based on a few characteristics provided
  - they also showed that with only 2 variables in a Logistic regression, similar results can be obtained...

Trade-of between fairness and accuracy

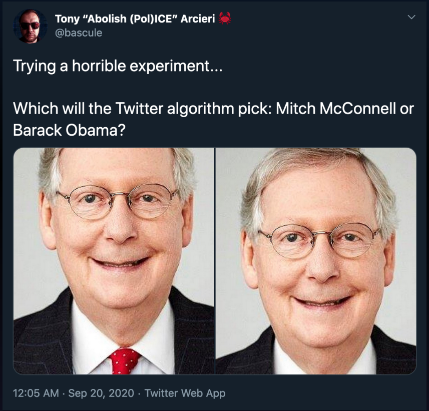
There is thus again a **trade-off between fairness** (demanded by society) and **algorithmic accuracy** (demanded by private sector) :

- Recall that the algorithm has an overall error rate of about 60%
- There is a high rate of false negative for White relative to Black people
- There is a high rate of false positive for Black relative to White people

If we would like false positive and false positive rates to be each relatively equal regardless of the “race” of the individuals (if we want more fairness), we can change the threshold value above which we consider the probability that a person will commit a further crime (this will degrade the overall error rate).

3.2.3 Transparency and fairness: when testing allows to spot biases

Racial bias in Twitter photo preview



Tony “Abolish (Pol)ICE” Arcieri 🦀
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

12:05 AM · Sep 20, 2020 · Twitter Web App

Figure 8: https://twitter.com/bascule/status/1307440596668182528

Racial bias in Twitter photo preview

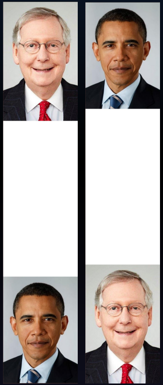
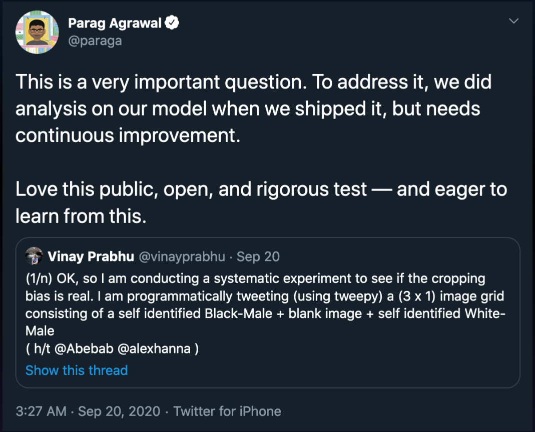


Figure 9: https://twitter.com/bascule/status/1307440596668182528

Once exposed, biased in the algorithms can be investigated



Parag Agrawal
@paraga

This is a very important question. To address it, we did analysis on our model when we shipped it, but needs continuous improvement.

Love this public, open, and rigorous test — and eager to learn from this.

Vinay Prabhu @vinayprabhu · Sep 20
(1/n) OK, so I am conducting a systematic experiment to see if the cropping bias is real. I am programmatically tweeting (using tweepy) a (3 x 1) image grid consisting of a self identified Black-Male + blank image + self identified White-Male
(h/t @Abebab @alexhanna)
Show this thread

3:27 AM · Sep 20, 2020 · Twitter for iPhone

Figure 10: https://twitter.com/paraga/status/1307491463639388160

3.3 Autonomy

Autonomy

- Using an algorithm to make a prediction and then to take decisions: **willingly ceding decision-making power to machines** (Floridi and Cowls 2019)
- Why do we do so? We believe that there is a **gain in efficiency** by turning to a machine learning model
- A trade-off between gain of efficacy and loss of control then arises.

Autonomy: two examples

Müller (2020) give two examples of autonomous systems and the problems raised:

1. Autonomous vehicles (AV)
    - Pros: should reduce the number of accidents (bodily injuring or not)
    - Cons: What should the decisions of the AV cars be? (as we saw in a previous section)
2. Autonomous weapons ("fire-and-forget" missiles)
    - Pros: **may** reduce war crimes and crimes in war, may reduce the number of deaths among the civilianscon
    - Cons: take responsability away of humans, increases the risks of war, makes killings more likely

Some additional reading

In this chapter, we talked about questions related to discriminations and fairness. There is a lot of research on those topics at the moment. The following by A. Charpentier report published in 2022 provides a great overview of these subjects: “Insurance: Discrimination, Biases & Fairness”

- English version
- French version

To finish

Ethics Guidelines for Trustworthy Artificial Intelligence (AI) (European Commission)

Figure 11: Darkside, Tom Stoppard https://vimeo.com/77882285

4. References

References I

Acemoglu, Daron, and Pascual Restrepo. 2019. "The Wrong Kind of AI? Artificial Intelligence and the Future of Labour Demand." *Cambridge Journal of Regions, Economy and Society* 13 (1): 25–35. https://doi.org/10.1093/cjres/rsz022.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica, May* 23: 2016.

Beam, Andrew L., and Isaac S. Kohane. 2018. "Big Data and Machine Learning in Health Care." *JAMA* 319 (13): 1317–18. https://doi.org/10.1001/jama.2017.18391.

Bonnefon, J.-F., A. Shariff, and I. Rahwan. 2016. "The Social Dilemma of Autonomous Vehicles." *Science* 352 (6293): 1573–76. https://doi.org/10.1126/science.aaf2654.

Cointe, Nicolas, Grégory Bonnet, and Olivier Boissier. 2016. "Ethical Judgment of Agents' Behaviors in Multi-Agent Systems." In *AAMAS*, 1106–14.

Davenport, Thomas, and Ravi Kalakota. 2019. "The Potential for Artificial Intelligence in Healthcare." *Future Healthcare Journal* 6 (2): 94–98. https://doi.org/10.7861/futurehosp.6-2-94.

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4 (1): eaao5580. https://doi.org/10.1126/sciadv.aao5580.

References II

Edwards, Lilian, and Michael Veale. 2018. "Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?" *IEEE Security & Privacy* 16 (3): 46–54. https://doi.org/10.1109/msp.2018.2701152.

Felzmann, Heike, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. "Transparency You Can Trust: Transparency Requirements for Artificial Intelligence Between Legal Norms and Contextual Concerns." *Big Data & Society* 6 (1): 205395171986054. https://doi.org/10.1177/2053951719860542.

Floridi, Luciano, and Josh Cowls. 2019. "A Unified Framework of Five Principles for AI in Society." *Issue 1*, June. https://doi.org/10.1162/99608f92.8cd550d1.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Kirkpatrick, Keith. 2015. "The Moral Challenges of Driverless Cars." *Communications of the ACM* 58 (8): 19–20. https://doi.org/10.1145/2788477.

Laat, Paul B. de. 2017. "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" *Philosophy & Technology* 31 (4): 525–41. https://doi.org/10.1007/s13347-017-0293-z.

References III

Lee, Michelle Seng Ah, and Luciano Floridi. 2020. "Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-Offs." *Minds and Machines*. https://doi.org/10.1007/s11023-020-09529-4.

Müller, Vincent C. 2020. "Ethics of Artificial Intelligence and Robotics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/; Metaphysics Research Lab, Stanford University.

O'neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Park, Andrew Lee. 2019. "Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing." *UCLA Lew Review*. https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/.

Piano, Samuele Lo. 2020. "Ethical Principles in Machine Learning and Artificial Intelligence: Cases from the Field and Possible Ways Forward." *Humanities and Social Sciences Communications* 7 (1). https://doi.org/10.1057/s41599-020-0501-9.

References IV

Rogers, David J., Sarah E. Randolph, Robert W. Snow, and Simon I. Hay. 2002. "Satellite Imagery in the Study and Forecast of Malaria." *Nature* 415 (6872): 710–15. https://doi.org/10.1038/415710a.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Settles, Burr, and Brendan Meeder. 2016. "A Trainable Spaced Repetition Model for Language Learning." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/p16-1174.

Sonnenburg, SÃķren, Mikio L Braun, Cheng Soon Ong, Samy Bengio, Leon Bottou, Geoffrey Holmes, Yann LeCun, et al. 2007. "The Need for Open Source Software in Machine Learning." *Journal of Machine Learning Research* 8 (Oct): 2443–66.

Taleb, Nassim Nicholas. 2007. *The Black Swan: The Impact of the Highly Improbable*. Vol. 2. Random house.

References V

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research." Office of the Secretary, United States.

Wong, Yuna, John Yurchak, Robert Button, Aaron Frank, Burgess Laird, Osonde Osoba, Randall Steeb, Benjamin Harris, and Sebastian Bae. 2020. *Deterrence in the Age of Thinking Machines*. RAND Corporation. https://doi.org/10.7249/rr2797.

Yu, Han, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. "Building Ethics into Artificial Intelligence." In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/779.

Zarsky, Tal Z. 2013. "Transparent Predictions." *U. Ill. L. Rev.*, 1503.