MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

AI, algorithmic decision making and Big Data: Risks and Opportunities

Francesca Lagioia

Giovanni Sartor

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

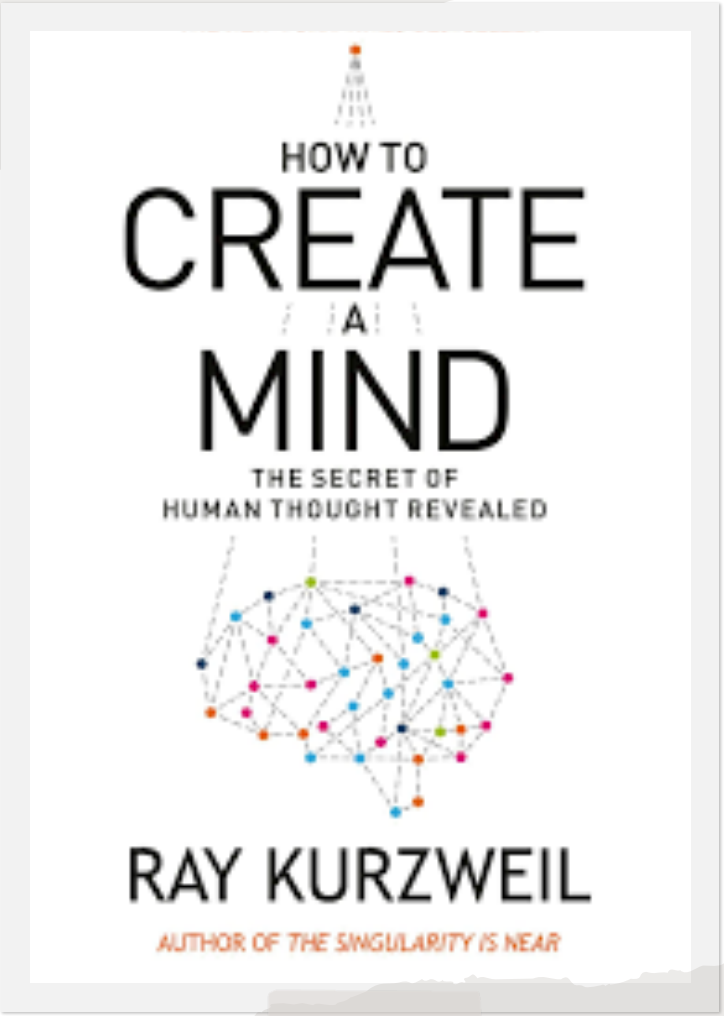
The Internet, AI and Big Data: promise and catch

The Internet & AI infrastructure can deliver good:

- It improves efficiency and effectiveness in many domains (smart cities, e-health, etc.)
- It allows for a world-wide generation and distribution of knowledge and solutions
- We can discover new correlations between things:
  - Doctors can provide better diagnoses and personalised and targeted therapies
- Cost savings, greater productivity, and value creation:
  - Firms can anticipate market trends and make more efficient decisions
  - Consumers can make better informed choices and obtain personalised services

AI opportunities: techno-optimistic perspective

Technologies based on artificial intelligence can allow humans to face the *"**the grand challenges of humanity, such as maintaining a healthy environment, providing the resources for a growing population (including energy, food, and water), overcoming disease, vastly extending human longevity, and eliminating poverty**"*. (Ray Kurzweil, *How to Create a Mind*)



AI and Big Data risks

- Eliminate or devalue the jobs of those who can be replaced by machines (exclusion and marginalization in the job market
- Lead to poverty and social exclusion
- Favour economic models in which *"the winner takes all".*
  - *Huge profits+limited workforce => AI contributes to concentrating wealth in those who invest in such companies or provide them with high-level expertise.*
- New opportunities for illegal activities
  - In particular, AI & Big Data systems can fall subject to cyberattacks (designed to disable critical infrastructure, or steal or rig vast data sets, etc.), and they can even be used to commit crimes (e.g., autonomous vehicles can be used for killing or terrorist attacks, and intelligent algorithms can be used for fraud or other financial crimes).

AI and Big Data risks

- Pervasive surveillance and manipulation
  - To satisfy data-hungry AI applications, the Internet has become an infrastructure for data collection (and surveillance)
  - All facts, even the apparently insignificant ones, are useful for learning algorithms, scalability is no problem

The power of AI can be used to purse **economic interests in ways that are harmful to individuals and society**: users, consumers, and workers can be subject to pervasive surveillance, controlled in their access to information and opportunities, manipulated in their choices.

Certain abuses may be incentivised by the fact that many tech companies —such as major platforms hosting user-generated content— operate in **two- or many-sided markets.**

AI and Big Data risks

- Their main services (search, social network management, access to content, etc.) are offered to individual consumers, but the revenue stream comes from advertisers, influencers, and opinion-makers (e.g., in political campaigns).
- This means not only that any information that is useful for targeted advertising will be collected and used for this purpose, but also that platforms will employ any means to capture users, so that they can be exposed to ads and attempts at persuasion.
- This may lead not only to a massive collection of personal data about individuals, to the detriment of privacy, but also to a pervasive influence on their behavior, to the detriment of both individual autonomy and collective interests.

AI and Big Data risks

- Polarization and fragmentation in the public sphere
  - proliferation of sensational and **fake news**, when used to capture users by exposing them to information they may like, or which accords with their preferences, thereby exploiting their confirmation biases .

- Just as AI can be misused by economic actors, it can also be **misused by the public section**. Governments have many opportunities to use AI for legitimate political and administrative purposes (e.g., efficiency, cost savings, improved services), but they may also employ it to anticipate and control citizens' behaviour in ways that restrict individual liberties and interfere with the democratic process.
- Restrict individual liberties and interfere with the democratic process
- Unfairness, discrimination and inequality

AI in decision making: approaches to learning

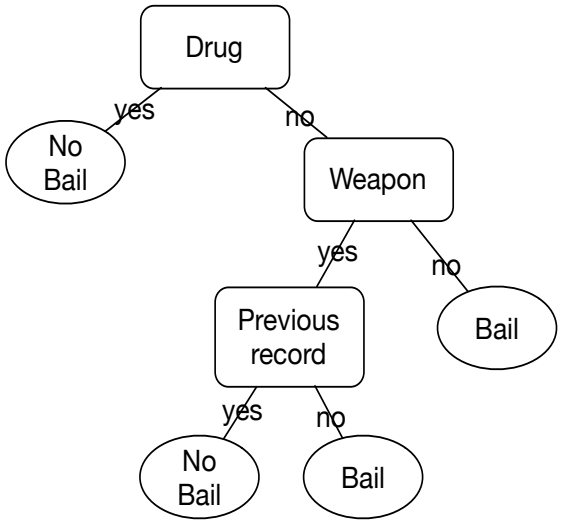
Supervised Learning	Unsupervised learning	Reinforcement learning
Machine is given examples of correct answers to cases It learns to answer in a similar way to new cases	Machine is given data It learns to identify patterns	Machine is given feedbacks (rewards and penalties) It learns by itself how to maximise its score

Supervised learning

- *Supervised learning* is currently the most popular approach. In this case the machine learns through "supervision" or "teaching":
- it is given in advance a training set, i.e., a large set of (probably) correct answers to the system's task. More exactly the system is provided with a set of pairs, each linking the description of a case to the correct response for that case.
- Here are some examples:
  - in systems designed to recognise objects (e.g. animals) in pictures, each picture in the training set is tagged with the name of the kind of object it contains (e.g., cat, dog, rabbit, etc);;
  - in systems for personnel selection, the description of each past applicants (age, experience, studies, etc.) is linked to whether the application was successful (or to an indicator of the work performance for appointed candidates);
  - in clinical decision support systems, each patient's symptoms and diagnostic tests is linked to the patient's pathologies;
  - in recommendation systems, each consumer's features and behaviour is linked to the purchased objects; in systems for assessing loan applications, each record of a previous application is linked to whether the application was accepted or not
- The training of a system does not always require a human teacher tasked with providing correct answers to the system. In many case, the training set can be side-product of human activities (purchasing, hiring, lending, tagging, etc.), as is obtained by recording the human choices pertaining to such activities
- In some cases the training set can even be gathered "from the wild" consisting in data which is available on the open web. (For instance, manually tagged images or faces, available on social networks)

Example of supervised learning: bail application

Predictors					Outcome
Case	**Injury**	**Drugs**	**Weapon**	**Prior-record**	**Decision**
1	none	no	no	yes	yes
2	bad	yes	yes	serious	no
3	none	no	yes	no	yes
4	bad	yes	no	yes	no
5	slight	yes	yes	yes	no
6	none	yes	yes	serious	no
7	none	no	yes	yes	no



- The decision tree captures the information in the training set through a combination of tests, to be performed sequentially. The first test concerns whether the defendant was involved in a drug related offence. If the answer is positive, we have reached the bottom of the tree with the conclusion that bail is denied. If the answer is negative, we move to the second test, on whether the defendant used a weapon, and so on.
- Notice that the decision tree does not include information concerning the kind of injury, since all outcomes can be explained without reference to that information. This shows how the system‘s model does not merely replicate the training set; it involves generalisation: it assumes that certain combination of predictors are sufficient to determine the outcomes, other predictors being irrelevant.

Predictions

The answers by learning systems are usually called “predictions”. However, often the context of the system’s use determines whether its proposals are be interpreted as forecasts, or rather as a suggestion to the system’s user.

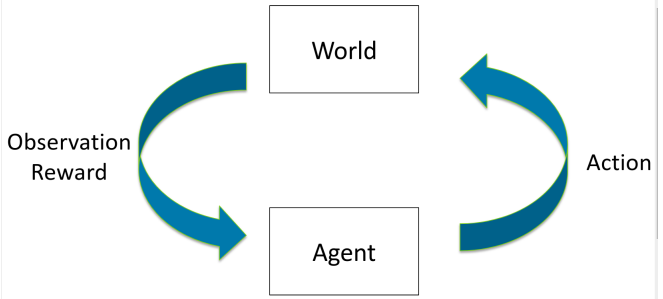
- For instance, a system’s “prediction” that a person’s application for bail or parole will be accepted can be viewed by the defendant (and his or her lawyer) as a prediction of what the judge will do, and by the judge as a suggestion guiding her decision (assuming that she prefers not to depart from previous practice).

Reinforcement learning

Reinforcement learning is similar to supervised learning, as both involve training by way of examples.

However, **in the case of reinforcement learning the systems learns from the outcomes of its own action, namely, through the rewards or penalties (e.g., points gained or lost) that are linked to the outcomes of such actions.**

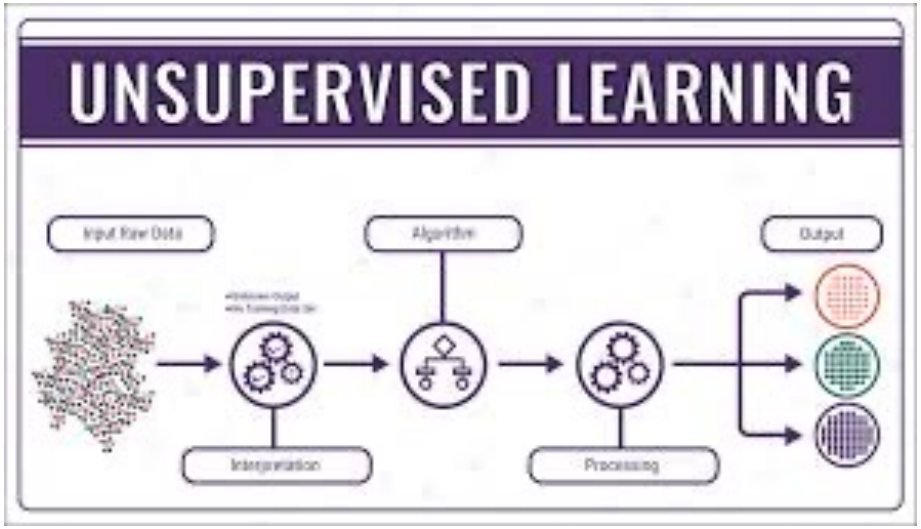
- E.g., in case of a system learning how to play a game, rewards may be linked to victories and penalties to defeats; in a system learning to make investments, rewards may be linked to financial gains and penalties to losses; in a system learning to target ads effectively, rewards may be linked to users' clicks, etc.



Unsupervised learning

In unsupervised learning, finally, AI systems learn without receiving external instructions, either in advance or as feedback, about what is right or wrong.

The techniques for unsupervised learning are used in particular, for **clustering**, i.e., **for grouping the set of items that present relevant similarities or connections** (e.g., documents that pertain to the same topic, people sharing relevant characteristics, or terms playing the same conceptual roles in texts).



For instance, in a set of cases concerning bail or parole, we may observe that injuries are usually connected with drugs (not with weapons as expected), or that people having prior record are those who are related to weapon. These clusters might turn out to be informative to ground bail or parole policies.

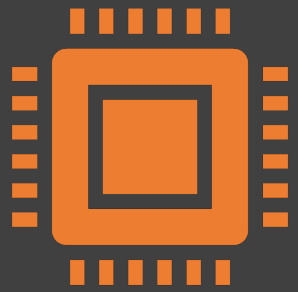
AI, Influence and Manipulation

Profiling, influence and manipulation

- The use of automated assessment systems may be problematic also where their performance is not worse, or even is better, than what humans would do.
- This is due to the fact that automation **diminishes the costs of collecting** information on individuals, **storing** this information and **process** it in order to **evaluate individuals and make choices accordingly.**
- Thus, automation paves the way for much more **persistent and pervasive mechanisms for assessment and control**.

- In general, thanks to AI, all kind of personal data can be used to **analyse**, **forecast** and **influence** human behaviour, an opportunity that transforms them into valuable commodities. Information that was not collected or was discarded as worthless “data exhaust” —e.g., trails of online activities—has now become a prized resource.

Profiling



Through AI & Big Data technologies—in combination with the panoply of sensor that increasingly trace any human activity—individuals can be subject to surveillance and influence in many more cases and contexts, on the basis of a broader set of personal characteristics (ranging from economic conditions to health situation, place of residence, personal life choices and events, online and offline behaviour, etc.).

By correlating data about individuals to corresponding classifications and predictions, AI increases the potential for *profiling*, namely, for inferring information about individuals or groups, and adopting assessments and decisions on that basis.

Profiling: the scenario

A profiling system establishes (predicts) that individuals having certain features F_1, also have a certain likelihood of possessing certain additional features F_2.

- For instance, assume that the system establishes (predicts) that those having a genetic patterns have the tendency to develop a higher than average chance to develop cancer, or that those having a certain education and job history or ethnicity have a certain higher-than-average likelihood to default of their debts). Then we may say that this system has profiled the group of the individuals possessing features F_1: it has added to the description (the profile) of these group a new segment, namely, the likelihood of possessing the additional features F_2.

Profiling: the scenario

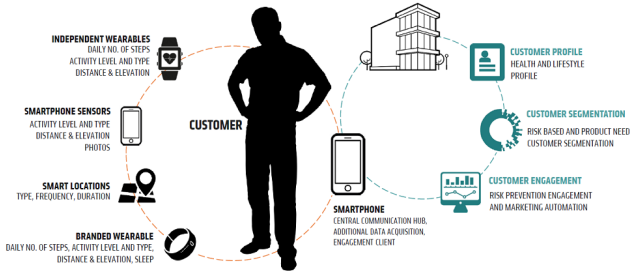
If the system is then given the information that a specific individual has features F_1, then the system can infer that it likely that this individual also has feature F_2. This may lead to the individual being treated accordingly, in a beneficial or a detrimental way.

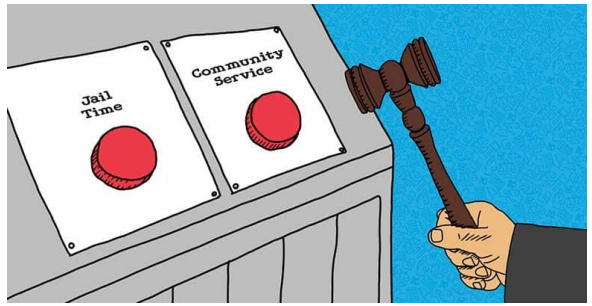
- For instance, in the case in which the inferred feature of an individual is his or her higher susceptibility to cancer, the system's indication may provide the basis for preventive therapies and tests, or rather for a raise in the insurance premium.

AI and profiling

- **AI & Big Data have vastly increased the opportunities for profiling.**
- Through the training, the system has learned an algorithmic model that can be applied to new cases: **if the model is given predictors-values concerning a new individual, it infers a corresponding target value for that individual, i.e., a new data item concerning him or her.**
  - ➢ the creditworthiness of loan applicants on the basis of their financial history, their online activity and social condition;
  - ➢ the likelihood that convicted persons may reoffend on the basis their criminal history, their character (as identified by personality test) and personal background.

These predictions may trigger automated determinations concerning, respectively, the price of a health insurance, the granting of a loan, or the release on parole.

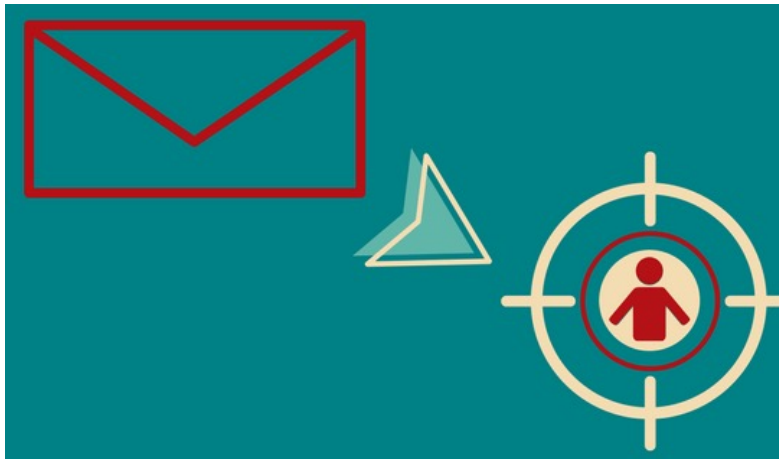




Profiling: influhence and manipulation

- **The information so inferred may also be conditional**, that is, it may consist in the **propensity to react in a certain way to given inputs.**
  - For instance, it may consist in in the propensity to respond to a therapy with improved medical condition, or in the propensity to respond to a certain kind of ad or to a certain price variation with a certain purchasing behaviour, or in the propensity to respond a certain kind of message with a change in mood or preference (e.g., relatively to political choices).
- **When that is the case, profiling potentially leads to influence and manipulation**.

Profiling: influhence and manipulation



- Assume, too, that the system connects certain values for input features (e.g., having a certain age, gender, social status, personality type, etc.) to the propensity to react to a certain message (e.g., a targeted ad) with a certain response (e.g., buying a certain product). Assume also that the system is told that a particular individual has these values (he is a young male, working class, extrovert, etc.).
- Then the system would know that by administering to the individual that message, the individual can probably be induced to deliver the response.

Even when an automated assessment and decision-making system —a profile-based system— **is unbiased, and meant to serve beneficial purposes, it may negatively affect the individuals concerned**. Those who are subject to pervasive surveillance, persistent assessments and insistent influence come under heavy psychological pressure that affects their personal autonomy, and they are susceptible to deception, manipulation and exploitation in multiple ways.

Profiling: a notion

- Profiling is a technique of (partly) automated processing of personal and/or non-personal data, aimed at producing knowledge by inferring correlations from data in the form of profiles that can subsequently be applied as a basis for decision-making.
- A profile is a set of correlated data that represents a (individual or collective) subject.
- Constructing profiles is the process of discovering unknown patterns between data in large data sets that can be used to create profiles.
- Applying profiles is the process of identifying and representing a specific individual or group as fitting a profile and of taking some form of decision based on this identification and representation.

Bosco et al (2015); see also Hildebrandt, M. (2009).

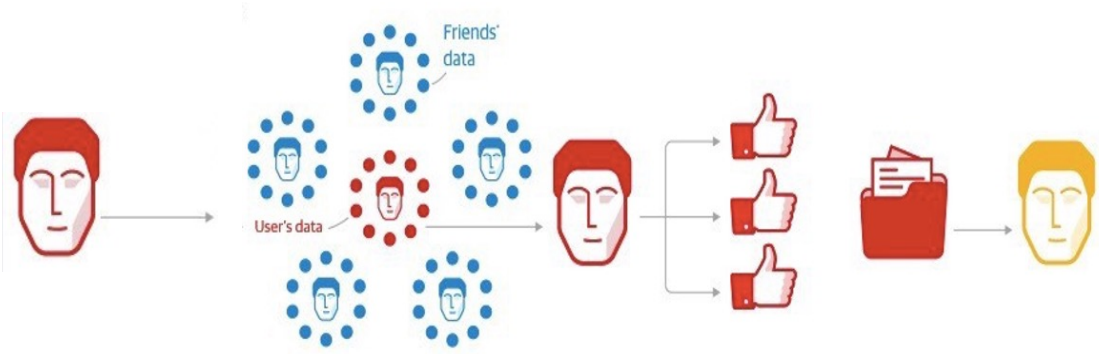
Profiling in GDPR

The notion of profiling in the GDPR only covers assessments or decisions concerning individuals, based on personal data, excluding the mere construction of group profiles:

- *'profiling'[…] consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces* ***legal effects*** *concerning him or her or* ***similarly significantly affects him or her.***

The dangers of profiling: the case of Cambridge Analytica

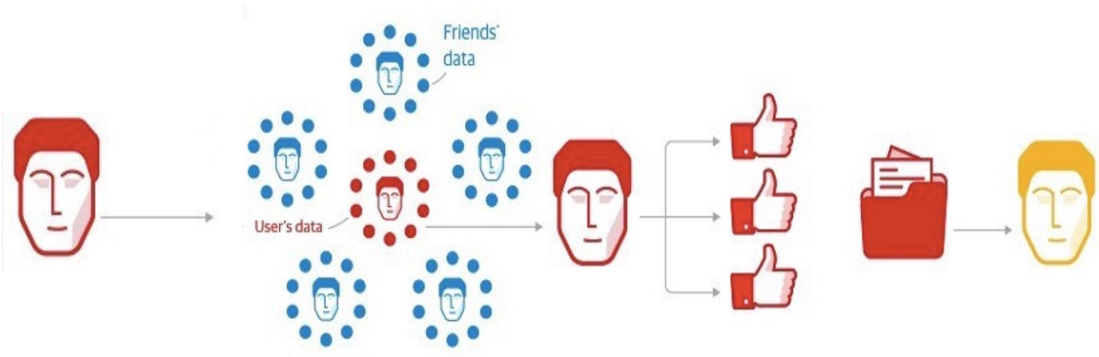
- First of all, people being registered as **voters** in the USA were invited to take a **detailed personality/political test** (about 120 questions), available online. The individuals taking the test would be **rewarded with a small amount of money** (from two to five dollars). They were told that their data would only be used for the academic research. About **320 000 voters took the test**. In order to receive the reward each individual taking the test had to **provide access to his or her Facebook page (step 1). This allowed the system to connect each individual's answers to the information included in his or her Facebook page.**



The dangers of profiling: the case of Cambridge Analytica

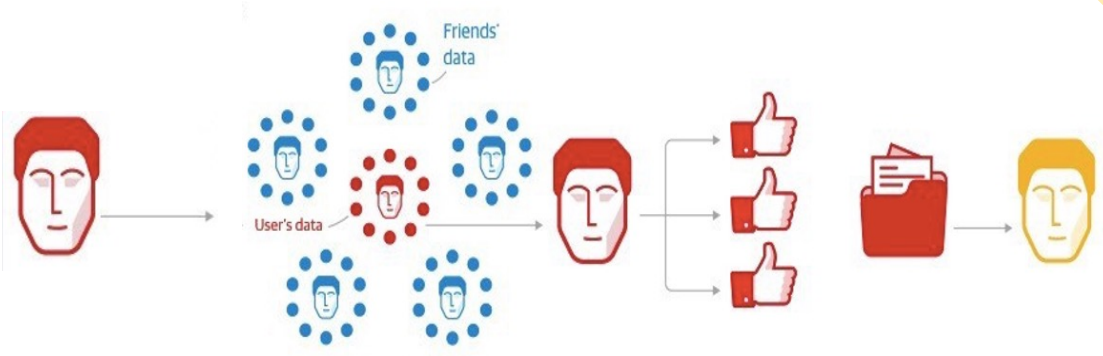
- When accessing a test taker's page, Cambridge Analytica **collected** not only **the Facebook page of test takers**, but also the Facebook **pages of their friends**, between 30 and 50 million people altogether **(step 2)**. Facebook data was also collected from other sources.

- After this data collection phase, Cambridge Analytica had at is disposition **two sets of personal data** to be processed **(step 3): (1) the data about the test takers**, consisting in the information on their **Facebook pages**, **paired with their answers** to the questionnaire, **(2)** and the **data about their friends**, consisting only in the information on their **Facebook pages.**



The dangers of profiling: the case of Cambridge Analytica

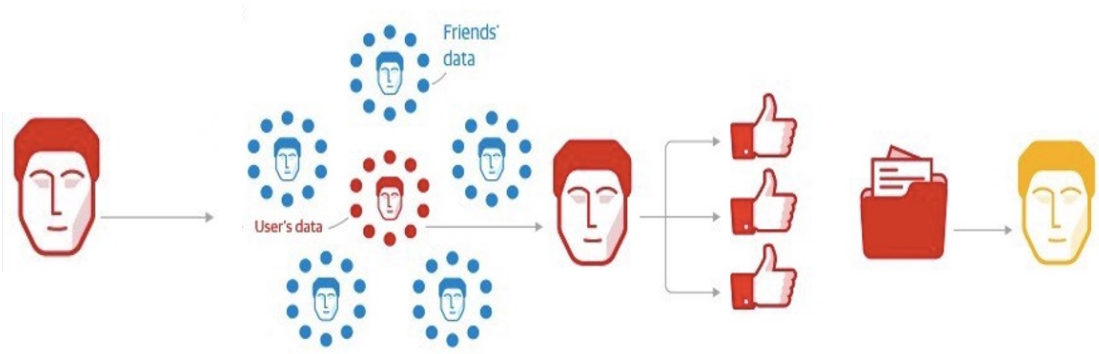
- Cambridge Analytica used the data about test-takers as a training set for building a model to profile their friends and other people. Data about the test-takers constituted a vast training set, where the information on an individual's **Facebook pages** (**likes, posts, links, etc.**) provided **values for predictors (features)** and **the answers to the questionnaire** (and psychological and political attitudes expressed by such answers) provided **values for the targets**. Thanks to its machine learning algorithms Cambridge Analytica could use this data to build a model **correlating the information in people's Facebook pages to predictions about psychology and political preferences.**



- Cambridge Analytica engaged in **massive profiling**, namely, in **expanding the data available on the people who did not take the test** (their Facebook data, and any further data that was available on them), **with the predictions provided by the model**. E.g. if test- takers having a certain pattern of Facebook likes and posts were classified as having a neurotic personality, the same assessment could be extended also to non-test-takers having similar patterns in their Facebook data.

The dangers of profiling: the case of Cambridge Analytica

Finally **(stage 4)**, based on this personality/political profiling, **potential voters who were likely to change their voting behaviour were identified** (initially **2m people in 11 US States** in which a small change could make a difference) if provided with appropriate messages. These **voters where targeted with personalised political ads** and with other messages that could trigger the desired change in voting behaviour, possibly building upon their emotions and prejudice and without making them aware of the purpose of such messages.



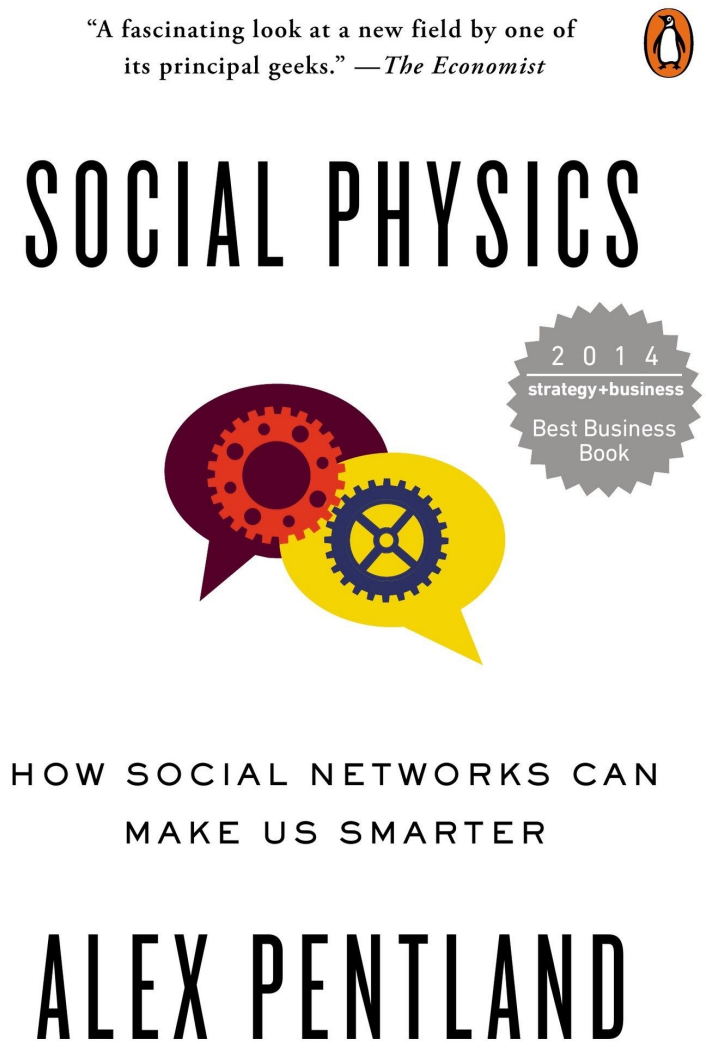
Towards surveillance capitalism or surveillance State?

- Some authors have taken a positive view of the development of systems based on the massive collection of information. They have observed that the integration of AI and Big Data enables **increased efficiency** and provides **new means for managing and controlling individual and social behaviour.**
- When economic transactions —and more generally social interaction and individual activities— are **computer-mediated**, they provide for a **ubiquitous and granular recording of data**: computer systems can observe, verify and analyse any aspects of the activities in question. **The recorded data can be used to construct user profiles, to personalise interactions with users** (as in targeted commercial communication), **to engage in experimentation** (e.g., to evaluate user responses to changes in prices and messaging), **to guide and control behaviour** (e.g., for the purpose of economic or political persuasion).

Towards surveillance capitalism or surveillance State?

In this context, **new models of economic and social interaction become possible, which are based on the possibility of observing every behaviour, and of automatically linking penalties and rewards to it.**

- Consider for instance how online consumers trust vendors of goods and services with whom they have never had any personal contact, relying on the platform through which such goods and services are provided, and on the platform's methods for **rating**, **scoring**, **selecting**, and **excluding**.



Towards surveillance capitalism or surveillance State?

According to Alex Pentland the director of the Human Dynamics Lab at the MIT Media Lab, AI & Big Data may enable the development of a **"social physics",** i.e., a rigorous social science. The availability of vast masses of data and of methods and computational resources to process these data could support a **social science having solid theoretical-mathematical foundations as well as operational capacities for social governance.**

Industrial Capitalism and Surveillance capitalism

The prospect for economic and social improvement offered by AI and Big Data is accompanied by the risks referred to as **"surveillance capitalism"** and the **"surveillance State".**

According to **Shoshana Zuboff**, **surveillance capitalism** is the **leading economic model** of the present age.

Industrial Capitalism

Karl Polanyi observed that industrial capitalism also treats as **commodities** (products to be sold in the market) **entities that are not produced for the market**: human life becomes "**labour**" to be bought and sold, nature becomes "**land**" or "**real estate**", exchange becomes "**money**."

As a consequence, the dynamics of capitalism produces destructive tensions —exploitation, destruction of environment, financial crises— unless **countervailing forces**, such as **law**, **politics** and **social organisations** (e.g., workers' and consumers' movements), intervene to counteract, moderate and mitigate excesses.

Industrial Capitalism and Surveillance capitalism

The prospect for economic and social improvement offered by AI and Big Data is accompanied by the risks referred to as **“surveillance capitalism”** and the **“surveillance State”.**

According to **Shoshana Zuboff**, **surveillance capitalism** is the **leading economic model** of the present age.

Surveillance Capitalism

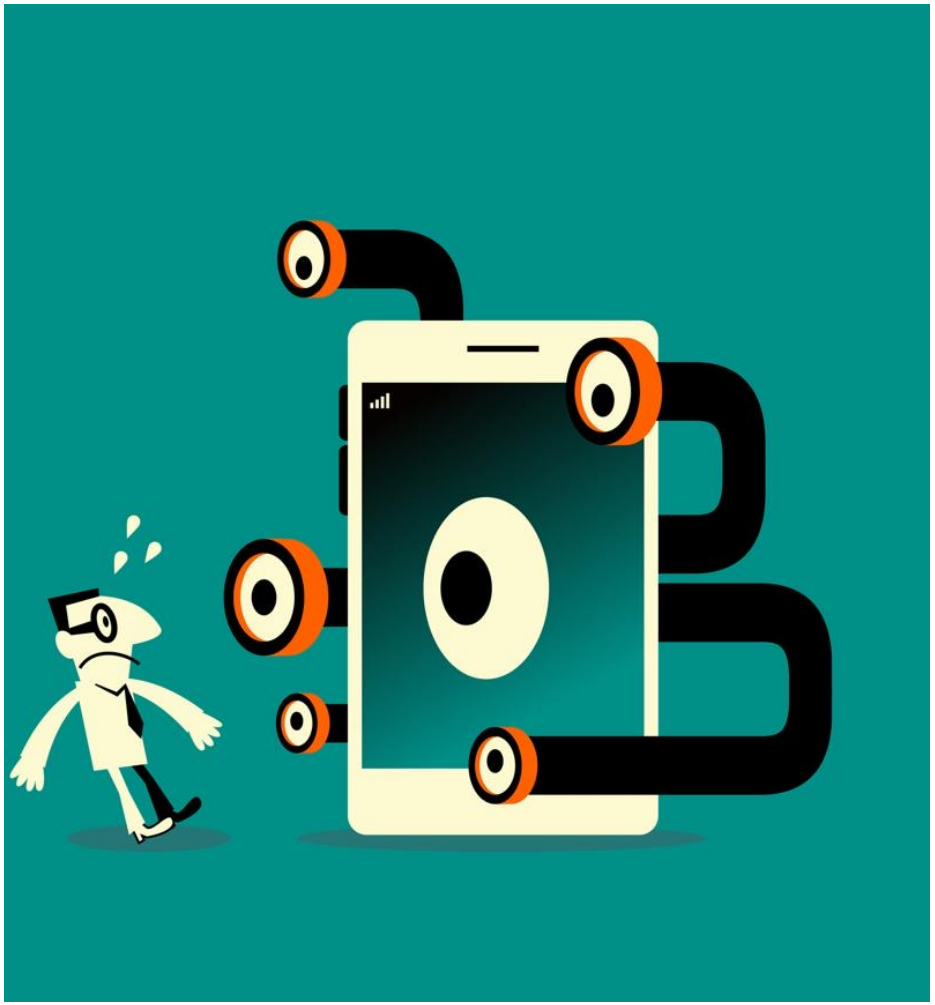
It expands commodification, extending it to **human experience**, which it turns into **recorded and analysed behaviour**, i.e., it transforms into **marketable opportunities to anticipate and influence**.
annexes **human experience** to the market dynamic so that it is reborn as behavior: **the fourth “fictional commodity.”**

Surveillance capitalism

Polanyi's first three **fictional commodities**—land, labor, and money—were subjected to **law**.

Although these laws have been imperfect, the institutions of labor law, environmental law, and banking law are regulatory frameworks intended to defend society (and nature, life, and exchange) from the worst excesses of raw capitalism's destructive power.

Surveillance capitalism's expropriation of human experience has faced **no** such **impediments**.



Surveillance capitalism

- In the case of surveillance capitalism, raw market dynamics can lead to novel disruptive outcomes. Individuals are subject to **manipulation**, are **deprived of control over their future** and **cannot develop their individuality.** Social networks for collaboration are replaced by surveillance-based mechanism of incentives and disincentives. (e.g. Uber recording workers'performance + mutual reviews of workers and clients)
- This new way of governing human behaviour may lead to efficient outcomes, but **it affects the mental wellbeing and autonomy of the individuals concerned.**

- According to Zuboff, we have not yet developed **adequate legal, political or social measures** by which to check the potentially disruptive outcomes of surveillance capitalism and keep them in balance. However, she observes, **the GDPR** could be an important step in this direction

- The need to **limit the commercial use of personal data** has led to new legal schemes not only in Europe, but also in California, the place where many world-leading "surveillance capitalists" have their roots; the CCPA (**California Consumer Privacy Act**), which came into effect on January 2020, provides consumers with rights to access their data and to prohibit data sales (broadly understood).

Surveillance State

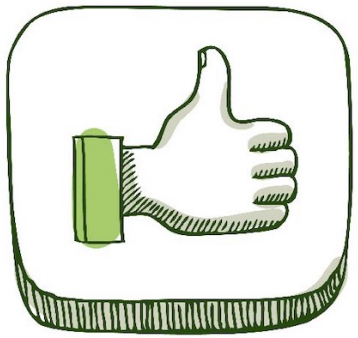
- At the governmental level, surveillance capitalism finds its parallel in the so-called “surveillance State”

- In the National Surveillance State, the government uses surveillance, data collection, collation, and analysis to **identify problems**, to **head off potential threats**, to **govern populations**, and to **deliver valuable social services**.

- The National Surveillance State is a special case of **the Information State**-a state that tries to identify and solve problems of governance through the collection, collation, analysis, and production of information.

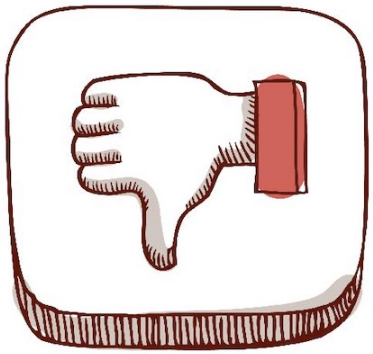
Surveillance State

Advantage



- Support efficiency in managing public activities
- Coordinate citizens' behaviour
- Prevent social harms

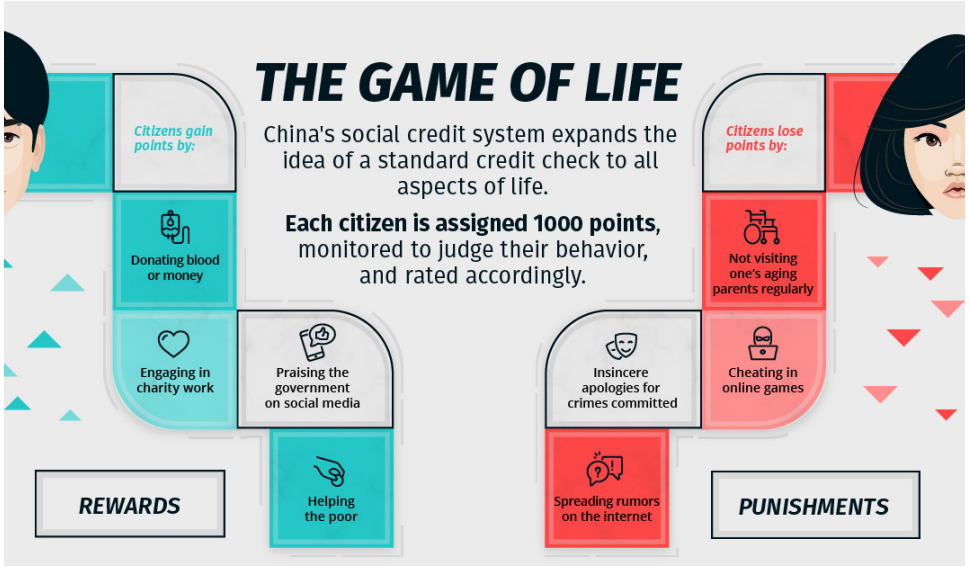
Disadvantage



- New kinds of influence and control
- Promote values and Purposes that may conflict with democracy
- Diminish autonomy

Surveillance State: the Chinese Social credit systems

- The Chinese Social credit systems collects data about citizens and assigns to those citizens scores that quantify their social value and reputation.
- It is based on the aggregation and analysis of personal information
- The collected data cover **financial aspects** (e.g., timely compliance with contractual obligations), **political engagement** (e.g., participation in political movements and demonstrations), **involvement in civil and criminal proceedings** (past and present) and **social action** (e.g. participation in social networks, interpersonal relationships, etc.).
- Citizens may be assigned positive or negative points, which contribute to their social score.



- The overall score determines citizens' **access to services and social opportunitie**s', e.g. universities, housing, transportation, jobs, financing, etc.
- The system's purported **objective** is to **promote mutual trust, and civic virtues.**
- **Risks**: **opportunism** and **conformism** may be rather promoted to the detriment of individual autonomy and genuine moral and social motivations.

Individual and social costs of AI & Big Data applications

In some cases and domain, AI & Big Data applications—even when accurate and unbiased—may have individual and social costs that outweigh their advantages.

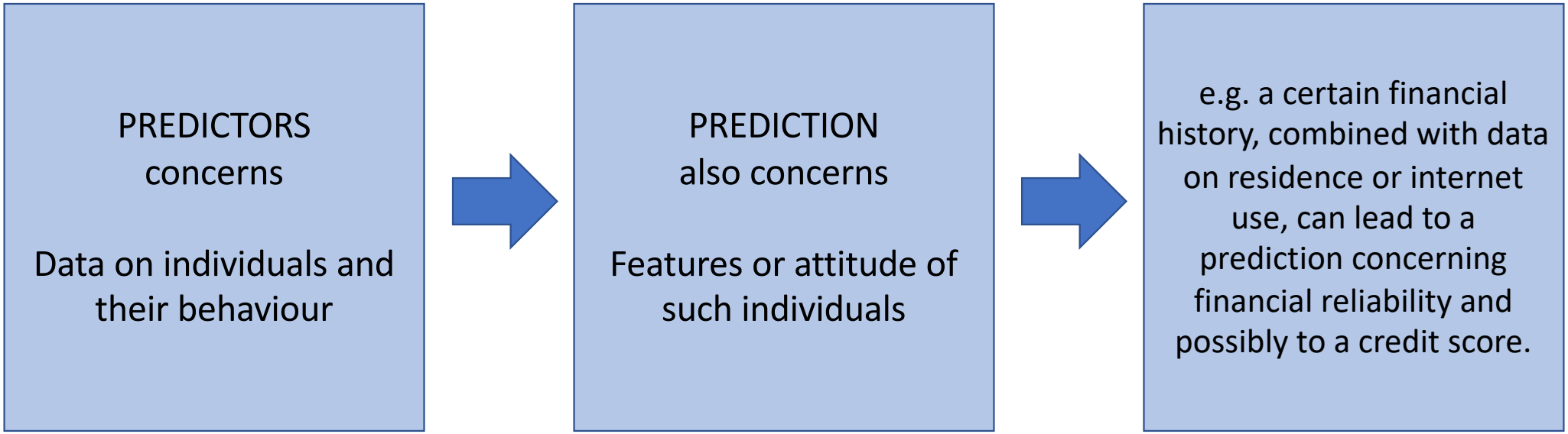
- Which systems really deserve to be built?
- Which problems most need to be tackled?
- Who is best placed to build them?
- And who decides?

We need genuine **accountability mechanisms**

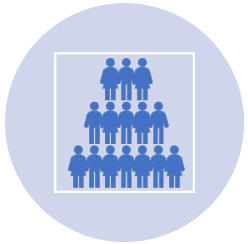
Consider, for instance, systems that are able to recognise sexual orientation, or criminal tendencies from the faces of persons. **Should we just ask whether these systems provide reliable assessments, or should we rather ask whether they should be built at all?**

The general problem of social sorting and differential treatment

The key aspect of ML systems is the ability to engage in **differential inference: different combinations of predictor-values are correlated to different predictions.**



The general problem of social sorting and differential treatment



A new dynamic of **stereotyping** and **differentiation** takes place.

(a) The individuals whose data support the same prediction, will be considered and treated in the same way.

(b) The individuals whose data support different predictions, will be considered and treated differently.

This **equalisation and differentiation**, depending on the domains in which it is used and on the purposes that it is meant to serve, may affect **positively or negatively** the individuals concerned.

Example: use of machine learning technologies to detect or anticipate health issues

- Beneficial application
- **Benefits concern in principle all data subjects** i.e.**,** those) whose data are processed for this purpose

- Processing of health-related data may also be justified on grounds of public health (Article 9 (2)(h)), and in particular for the purpose of "monitoring epidemics and their spread" (Recital 46).
- This provision has become hugely relevant in the context of the Coronavirus disease 2019 (COVID-19) epidemics. In particular a vast debate has been raised by development **of applications for tracing contacts.**

Example: use of machine learning technologies to detect or anticipate health issues

- Such processing should be viewed as **legitimate** as long as it effectively contributes to **limit the diffusion and the harmfulness of the epidemics**, assuming that the **privacy and data protection risks are proportionate to the expected benefit**, and that **appropriate mitigation measures are applied**.

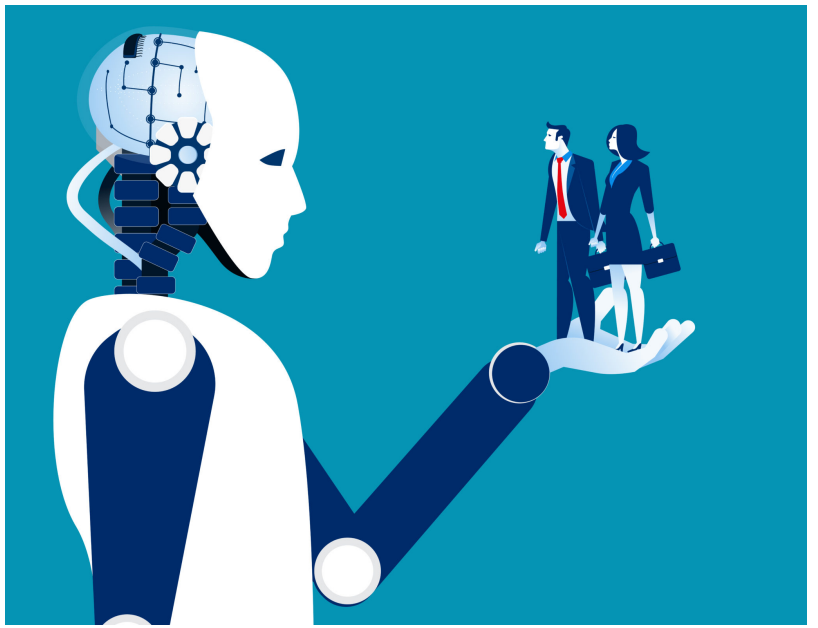
(See the European Data Protection Board Guidelines 04/2020 on the use of location data and contact-tracing tools in the context of the COVID-19 outbreak).

Example: use of the predictions based on health data in the contexts of insurance and recruting

- Predictions based on health data in the context of insurance deserves a **much less favourable assessment**
- **Gainers:** the insured individuals getting a better deal based on their favourable heath prospects.
- **Losers:** those getting a worse deal because of their unfavourable prospects.
- Individuals who already are disadvantaged because of their medical conditions would suffer further disadvantage, being excluded from insurance or being subject to less favourable conditions.

Example: use of the predictions based on health data in the contexts of insurance and recruting

- Insurance companies having the ability to distinguish the risks concerning different applicants would have a **competitive advantage**, being able to provide better conditions to less risky applicants, so that insurers would be pressured to collect as much personal data as possible.
- **Even less commendable would be the use of health predictions in the context of recruiting**, which would involve burdening less healthy people with unemployment or with harsher work conditions. Competition between companies would also be affected, and pressure for collecting health data would grow.



A further example concerns…

Price discrimination: different prices and different conditions to different consumers, depending on predictions on their willingness to pay.

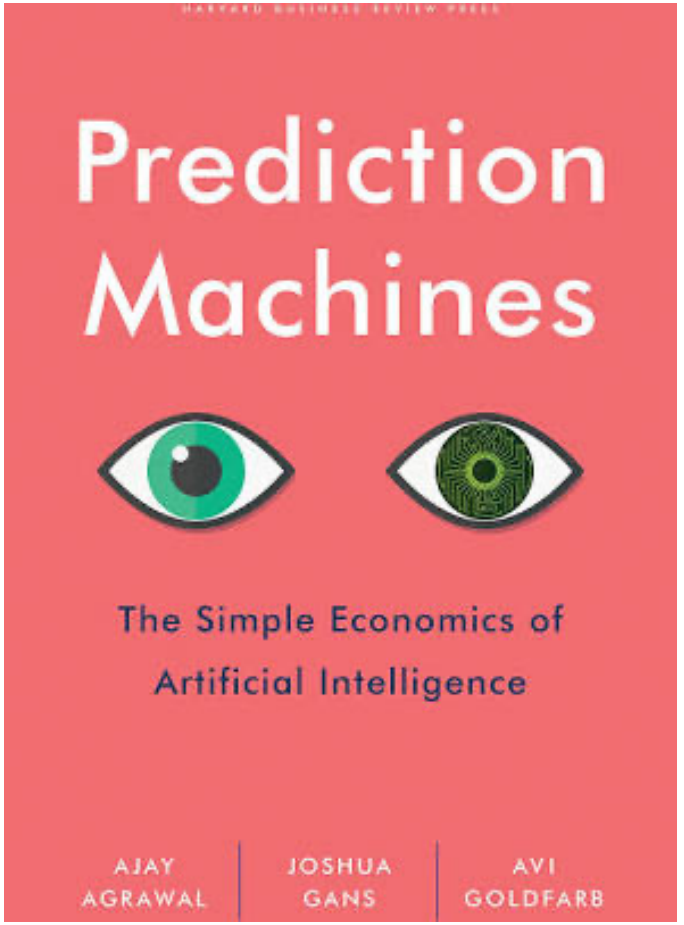
Risks:

- harm consumers
- individuals may be deprived of access to some opportunities
- affect the functioning of markets
- may be unfair(?)
- undermines the efficiency of the economy

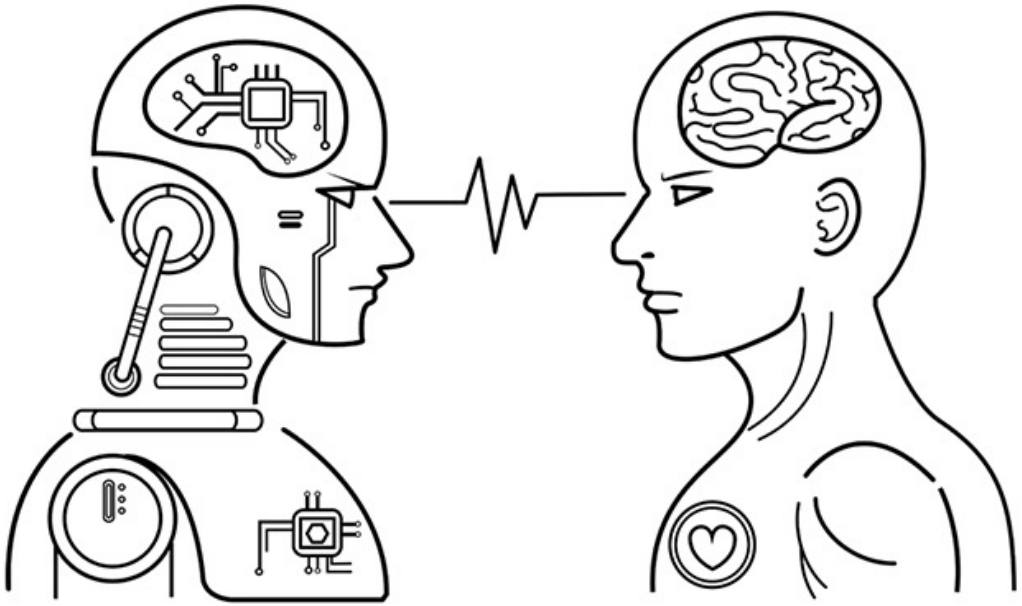
Example: price discrimination

AI in decision making concerning individuals: fairness and discrimination

- The combination of AI and Big Data enables automated decision-making even in domains that require complex choices, based on multiple factors, and on non-predefined criteria.
- In recent years, a wide debate has taken place on prospects and risks of algorithmic assessments and decisions concerning individuals



Are AI systems better then humans in assessing us?



In many domains automated predictions and decisions are not only **cheaper**, but also **more precise and impartial** than human ones.

- AI can **avoid** typical **fallacies of human psychology** (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc,), and the widespread human **inability to process statistical data,** as well as **typical human prejudice** (concerning, e.g., ethnicity, gender, or social background).
- In many assessments and decisions —on investments, recruitment, creditworthiness, or also on judicial matters, such as bail, parole, and recidivism— algorithmic systems have **often performed bette**r, according to usual standards, than human experts.

Or not?

Others have underscored the possibility that algorithmic decisions may be **mistaken or discriminatory.**

- Only in rare cases will algorithms engage in explicit unlawful discrimination, so-called **disparate treatment**, basing their outcomes on prohibited features (predictors) such as race, ethnicity or gender.
- More often a system's outcome will be discriminatory due to its **disparate impact**, i.e., since it disproportionately affects certain groups, without an acceptable rationale

Systems reproducing the strengths and weaknesses of humans in making judgments

Systems based on **supervised learning** may be trained on **past human judgements** and may therefore **reproduce** the strengths and weaknesses of the humans who made these judgements, including their **propensities to error and prejudice.**

- For example, a recruitment system trained on the past hiring decisions will learn to emulate the managers' assessment of the suitability of candidates, rather than to directly predict an applicant's performance at work. If past decisions were influenced by prejudice, the system will reproduce the same logic.

Prejudice in the training set

Prejudice baked into training sets may persist even if the inputs (the predictors) to automated systems do not include forbidden discriminatory features (e.g. ethnicity or gender.)

This may happen whenever **a correlation exists between discriminatory features and some predictors**

- Assume, for instance, that a prejudiced human resources manager did not hire applicants from a certain ethnic background, and that people with that background mostly live in certain neighbourhoods. A training set of decisions by that manager will teach the systems not to select people from those neighbourhoods, which would entail continuing to reject applications from the discriminated-against ethnicity. (Kleinberg et all (2019)).

Systems biased against groups

In other cases, a training set may be biased against a certain group, since the achievement of the outcome being predicted (e.g., job performance) is approximated through a **proxy** that has a disparate impact on that group.

- Assume, for instance, that the **future performance** of employees (the target of interest in job hiring) is only measured by the **number of hours worked in the office.** This outcome criterion will lead to past hiring of women —who usually work for fewer hours than men, having to cope with family burdens— being considered less successful than the hiring of men; based on this correlation (as measured on the basis of the biased proxy), the systems will predict a poorer performance of female applicants.

System's biases embedded in the predictors

In other cases, mistakes and discriminations may pertain to the machine-learning system's biases embedded in the predictors.

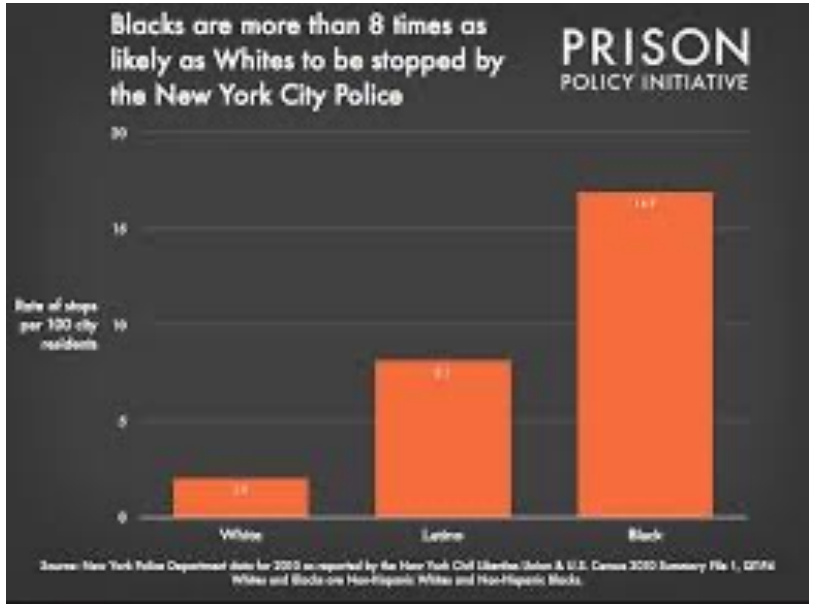
A system may perform unfairly, since it uses a favourable predictor (input feature) that only applies to members of a certain group (e.g., the fact of having attended a socially selective high-education institution).

Unfairness may also result from taking biased human judgements as predictors (e.g., recommendation letters).

Data set that does NOT reflect the statistical composition of the population

Finally, unfairness may derive from **a data set that does reflect the statistical composition of the population.**

- Assume for instance that in applications for bail or parole, previous criminal record plays a role, and that members of a certain groups are subject to stricter controls, so that their criminal activity is more often detected and acted upon. This would entail that members of that group will generally receive a less favourable assessment than members of other groups having behaved in the same ways.



- Members of a certain group may also suffer prejudice when that group is only represented by a very small subset of the training set,
- This will reduce the accuracy of predictions for that group (e.g., consider the case of a firm that has appointed few women in the past and which uses its records of past hiring as its training set).

Challenging the unfairness of automated decision-making

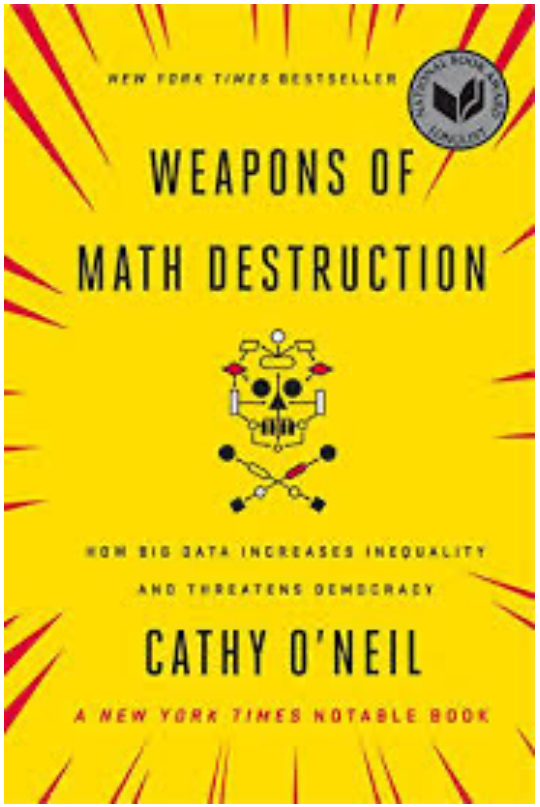
It has been observed that it is difficult to challenge the unfairness of automated decision-making.

Challenges raised by the individuals concerned, even when justified, may be disregarded or rejected because they interfere with the system's operation, giving rise to additional **costs and uncertainties**.

In fact, predictions of machine-learning systems are based on **statistical correlations**, **against which it may be difficult to argue** on the basis of individual circumstances, even when exceptions would be justified.

Weapons of math distruction

“An algorithm processes a slew of statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone’s life upside down. And yet when the person fights back, “suggestive” countervailing evidence simply won’t cut it. The case must be ironclad. The human victims of WMDs, we’ll see time and again, are held to a far higher standard of evidence than the algorithms themselves”. (O’Neil (2016))



Or not?

[W]ith appropriate requirements in place, the use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred. By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central **trade-offs among competing values.** Algorithms are not only a threat to be regulated; with the right safeguards in place, they have **the potential to be a positive force for equity**

(Kleinberg, Ludwig, Mullainathan, e Sunstein (2018, 113)).

Challenging the unfairness of automated decision-making

These criticisms have been countered by observing that **algorithmic systems,** even when based on machine learning, are **more controllable** than human decision-makers, their **faults can be identified** with precision, and **they can be improved** and **engineered to prevent unfair outcomes**.

Should we exclude the use of automated decision-making?

It seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making.

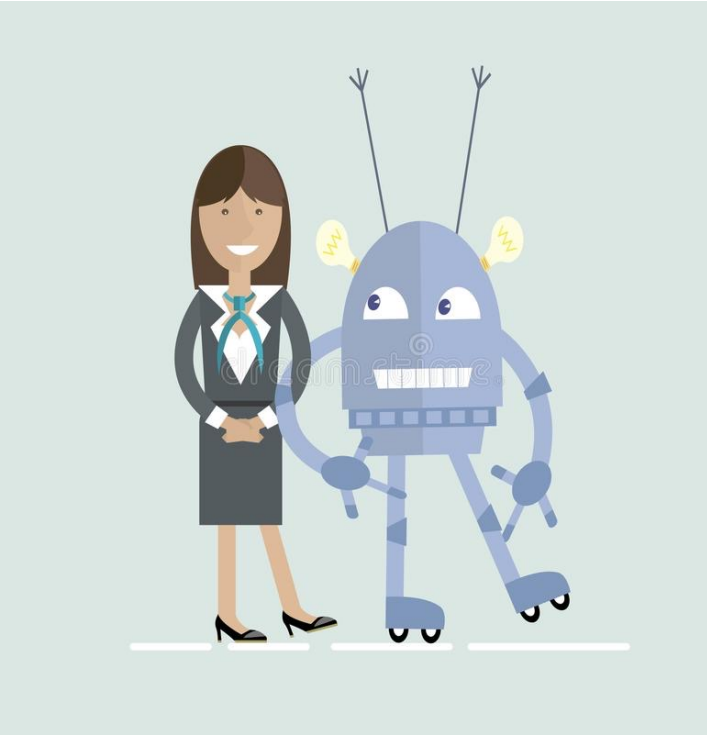
The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker.

Humans + Algorithms?

In many cases, the best solution consists in **integrating human and automated judgements**, by enabling the affected individuals to request a **human review** of an automated decision as well as by favouring **transparency** and developing methods and technologies that enable human experts to analyse and review automated decision-making.

In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, e.g., medical diagnosis, financial investment, granting of loans, etc.

The future challenge will consist in finding the best combination between human and AI, taking into account the capacities and the limitations of both.



Conclusions..

- AI enables new kinds of algorithmic mediated differentiations between individuals
- In the **AI era differential treatments** can be based on **vast amounts of data enabling probabilistic predictions, which may trigger algorithmically predetermined responses**.
- The impacts of such practices can go beyond the individuals concerned, and affect important social institution, in the economical and political sphere.

Conclusions..

The **GDPR** provides some **constraints**:

- the need for a legal basis for any processing of personal data
- obligations concerning information and transparency
- limitations on profiling and automated decision making
- requirements on anonymisation and pseudonymisation, etc.

These constraints need to be coupled with **strong public oversight**, to ban socially obnoxious forms of differential treatment, and to adopt effective measures that prevent abuses.

Some interests at stake

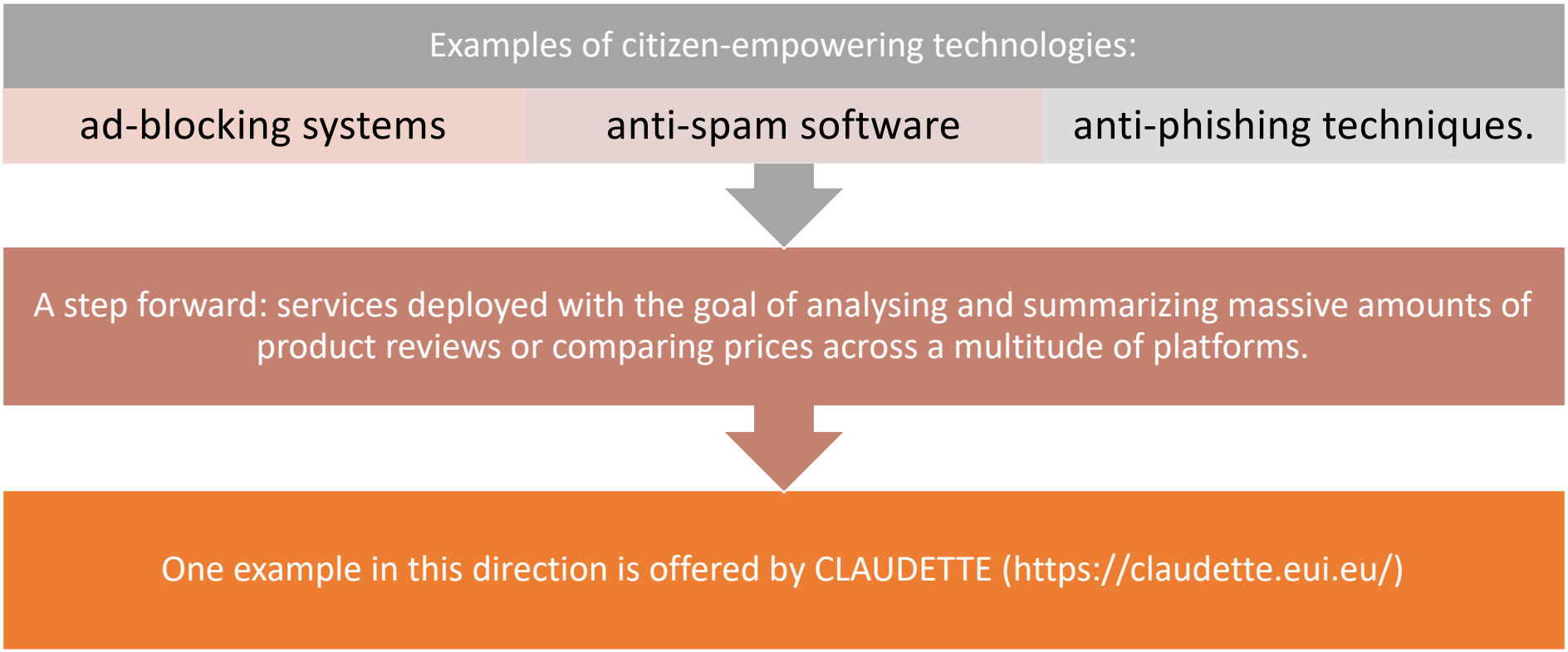
- **Interest in data protection and privacy,** namely, the interest in a lawful and proportionate processing of personal data subject to oversight.
- **Interest in fair algorithmic treatment**: concern from an algorithmic transparency/explicability standpoint
- **Individual autonomy:** black boxe models whose functioning is not accessible and whose decisions remain unexplained and thus unchallengeable.
- **Interest in not being misled or manipulated by AI systems and to trust** such systems.
- **Indirect interest in fair algorithmic competition**, i.e., in not being subject to market-power abuses resulting from exclusive control over masses of data and technologies.

AI technologies for social and legal empowerment

Regulatory instruments and their implementation by public bodies are an essential element but they may be insufficient.

- AI & Big Data are employed in domains already characterized by a **vast power imbalance**, which they may contribute to accentuate.
- The **countervailing power of civil society is needed** to detect abuses, inform the public, activate enforcement, etc.
- In the AI era, an effective **countervailing power needs also to be supported by AI**

Citizen-empowering technologies - Claudette



Citizen-empowering technologies – PDA/CDA

- Proposals for automatically extracting, categorizing and summarizing information from **privacy documents**, and assisting users in processing and understanding their contents.

- Multiple AI methods to support data protection could be merged into integrated PDA-CDA **(Privacy digital assistants/consumer digital assistants),** meant to prevent excessive/unwanted/unlawful collection of personal data and well as to protect users from manipulation and fraud, provide them with awareness of fake and untrustworthy information, and facilitate their escape from "**filter bubbles" (the unwanted filtering/pushing of information).**

MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423