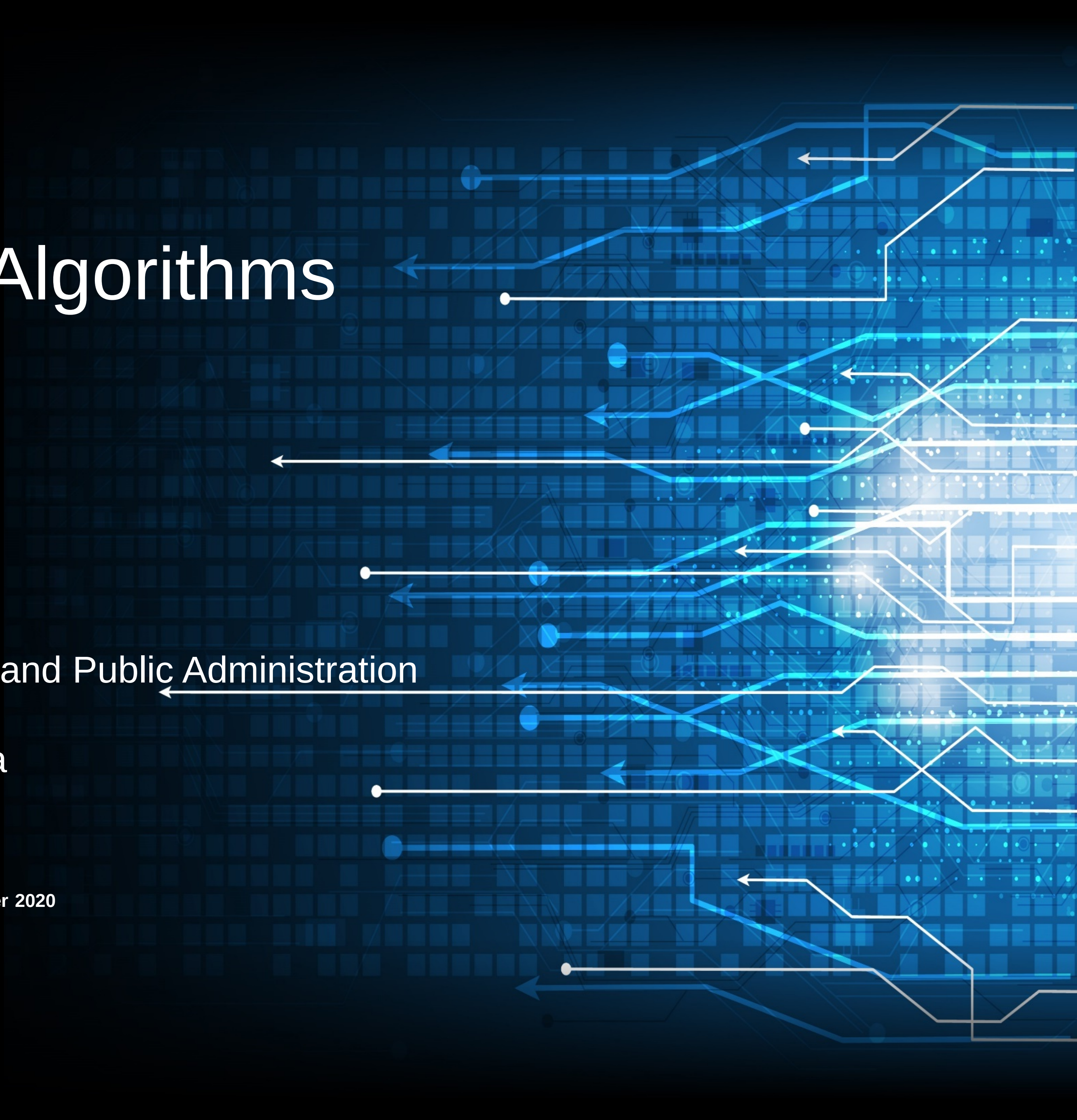
punt de trobada

ApLAB

Approval of Algorithms



Ministry of Digital Policy and Public Administration

Generalitat de Catalunya

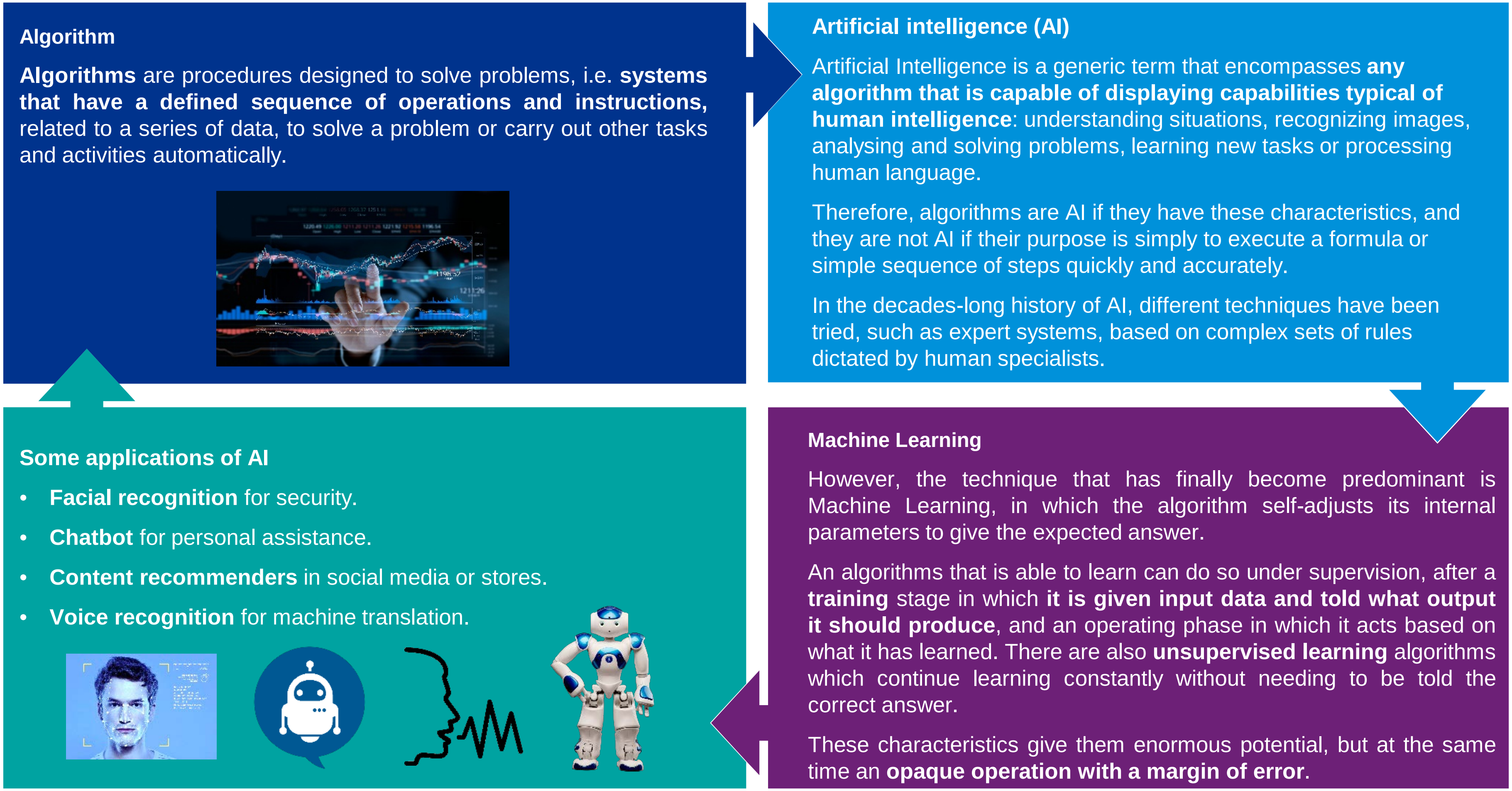
Dossier for the ApLab session of 10 November 2020

Contents

1. Algorithms

Basic concepts

Algorithms are increasingly present in the day-to-day lives of citizens, and specifically in the public sector. The following diagram shows several algorithmic concepts, linked by the concepts of algorithm and artificial intelligence.



Algorithm

Algorithms are procedures designed to solve problems, i.e. **systems that have a defined sequence of operations and instructions,** related to a series of data, to solve a problem or carry out other tasks and activities automatically.

Artificial intelligence (AI)

Artificial Intelligence is a generic term that encompasses **any algorithm that is capable of displaying capabilities typical of human intelligence**: understanding situations, recognizing images, analysing and solving problems, learning new tasks or processing human language.

Therefore, algorithms are AI if they have these characteristics, and they are not AI if their purpose is simply to execute a formula or simple sequence of steps quickly and accurately.

In the decades-long history of AI, different techniques have been tried, such as expert systems, based on complex sets of rules dictated by human specialists.

Machine Learning

However, the technique that has finally become predominant is Machine Learning, in which the algorithm self-adjusts its internal parameters to give the expected answer.

An algorithms that is able to learn can do so under supervision, after a **training** stage in which **it is given input data and told what output it should produce**, and an operating phase in which it acts based on what it has learned. There are also **unsupervised learning** algorithms which continue learning constantly without needing to be told the correct answer.

These characteristics give them enormous potential, but at the same time an **opaque operation with a margin of error**.

Some applications of AI

- **Facial recognition** for security.
- **Chatbot** for personal assistance.
- **Content recommenders** in social media or stores.
- **Voice recognition** for machine translation.

1. Algorithms

Risks

The use of artificial intelligence algorithms on a large scale means that legislation and control must also adapt and grow to the same extent. We can classify the challenges of artificial intelligence in the public sector as follows:

Inequality

- AI algorithms learn and reproduce the inequalities present in the data with which they have been trained.
- Sometimes these inequalities are present in society.
- Some algorithms are more accessible to certain groups and can therefore favour them.

Algorithms may behave differently for different groups. It is necessary to review these behaviours and define which ones are unfair.

Lack of reliability

- Security and vulnerability issues with new automated AI systems.
- Staff with little training on how to identify whether the results of the algorithm are correct or not, or to understand why the algorithm produces the results it does.
- Possible errors during the execution of the algorithm.

AI algorithms do not follow exact rules, but have a probabilistic basis; the reliability must be assessed in each case.

Opacity

- When the decision-making criteria and processes affect people’s rights, an explication of them must be given.
- AI systems have a low level of transparency, both in their general criteria and in the explanation of each specific decision.

The public must not be left defenceless, rather they must be able to question the decisions made by algorithms.

Privacy and freedoms

- Possible effects on personal data protection.
- Technology offers benefits in exchange for privacy, and alters the balance of power between people, institutions and companies; where should we draw the line?

Several countries and large corporations are halting facial recognition projects.

2. Tools for algorithm harm prevention

The risks involved in the application of algorithms in public and private tasks have led to the creation around the world of a very diverse range of preventive and corrective instruments. The following are some examples.

Regulation of automated administrative tasks

Law 40/2015 on the Public Sector Legal System defines an automated administrative action as “any action or task carried out entirely by electronic means by a Public Administration within the framework of an administrative procedure and in which a public employee has not directly intervened”. In this case, the Administration must first establish the competent body or bodies:

- responsible for defining the specifications, programming, maintenance, oversight and quality control and, where appropriate, auditing the information system and its source code;
- against which appeals can be brought.

Automated administrative actions can use electronic signature systems based on the seal of the body or a secure verification code.

Furthermore, the European General Data Protection Regulation establishes that “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”, but this shall not apply if the decision:

- is necessary for entering into, or performance of, a contract;
- is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests; or
- is based on the data subject’s explicit consent.

Therefore, this regulation has no effect in these other cases:

- Algorithms used with the mediation of a public employee, who takes into account the recommendations of the algorithm but personally decides on the administrative task.
- Algorithms applied in the public sector outside the framework of an administrative procedure, such as, for example, a public service chatbot.

2. Tools for algorithm harm prevention

Awareness

One way of avoiding the harmful effects of algorithms is to raise awareness within the institutions that use them, and statements of principles are one way to do this. This is the case of the **Universal Guidelines for Artificial Intelligence** drawn up by the coalition of international organizations The Public Voice, which promotes public participation in decisions related to the future of the Internet. These guidelines aim to maximize the benefits of AI, minimize its risk and ensure the protection of human rights.

RIGHT TO TRANSPARENCY
All individuals have the right to know the basis of an AI decision that concerns them.

RIGHT TO HUMAN DETERMINATION
All individuals have the right to a final determination made by a person.

IDENTIFICATION
The institution responsible for an AI system must be made known to the public.

FAIRNESS
Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.

ASSESSMENT AND ACCOUNTABILITY
An AI system should be deployed only after an adequate evaluation of its purpose and objectives.

ACCURACY, RELIABILITY AND VALIDITY
Institutions must ensure the accuracy, reliability and validity of decisions.

DATA QUALITY
Institutions must establish data provenance and assure quality and relevance for the data input into algorithms.

PUBLIC SAFETY
Institutions must assess the public safety risks that arise from the deployment of AI systems.

CYBERSECURITY
Institutions must secure AI systems against cybersecurity threats.

PROHIBITION ON SECRET PROFILING
No institution shall establish or maintain a secret profiling system.

PROHIBITION ON UNITARY SCORING
No national government shall establish or maintain a general-purpose score on its citizens or residents.

TERMINATION OBLIGATION
An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.

The Allegheny County AI User Training Programme

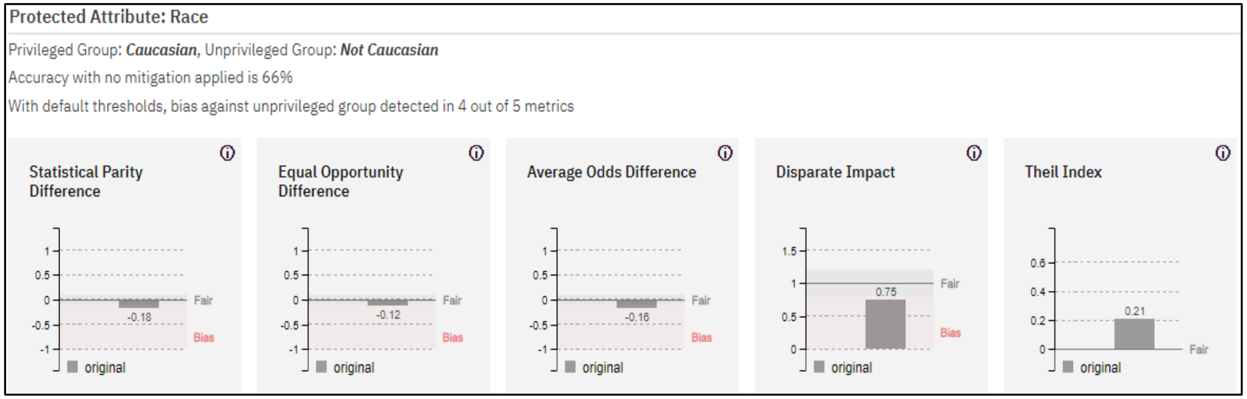
The **Allegheny Family Screening Tool** is an algorithm that helps Social Services identify children who should be removed from their families to avoid abuse. The algorithm suffers from a bias because it is based on public health data, meaning it is easier for wealthy families to hide abuse as they have access to private health services.

Staff working with this algorithm have received **specific training** which has allowed them to identify possible cases where data has been hidden and to become more aware of the issue. Through this training programme they were able to eliminate the bias produced by the algorithm.

2. Tools for algorithm harm prevention

Bias detection and mitigation

The performance of self-learning AI algorithms is conditioned by the fairness or bias of the data used to train them. There are several statistical tools that enable the identification of biases in training data. These tools can be found packaged for ease of use in toolkits such as AI Fairness 360º, released by IBM in 2018 as the first of its kind. This toolkit can be tested with any dataset that is loaded, or alternatively with the datasets of algorithms involved in famous cases of algorithm discrimination.



Protected Attribute: Race

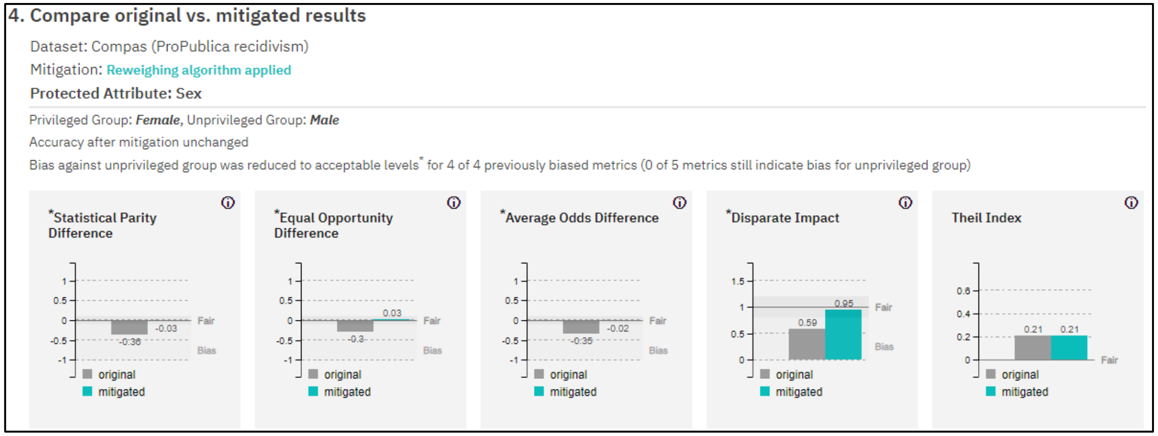
Privileged Group: *Caucasian*, Unprivileged Group: *Not Caucasian*

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

The application, which can be used online or by downloading the source code, uses five metrics to determine whether a dataset is fair or unfair with respect to various characteristics that the user wants to protect in each case, such as sex, race or age.

Biases in the input data can be corrected by obtaining fairer real data, or by means of workarounds such as the generation of artificial data or the reweighting of the various groups. The objective of toolkits like AI Fairness 360º is not to distinguish between good and bad algorithms but to help the institutions that create them to make them fairer. The application makes it possible to apply various bias mitigation techniques and see what improvements they bring in the five previously measured fairness metrics.



4. Compare original vs. mitigated results

Dataset: Compas (ProPublica recidivism)

Mitigation: Reweighing algorithm applied

Protected Attribute: Sex

Privileged Group: *Female*, Unprivileged Group: *Male*

Accuracy after mitigation unchanged

Bias against unprivileged group was reduced to acceptable levels* for 4 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group)

2. Tools for algorithm harm prevention

Transparency: open source

One of the principles of the ethical application of AI is that of transparency: people have the right to know the basis of an AI decision that affects them, a guideline that is equivalent to the administrative law principle that states that administrative decisions must be justified, i.e. the reasons for the decision adopted must be expressed rationally.

Algorithms are implemented using computer programs that can have thousands of lines of code. One of the approaches to achieving transparency is to establish the obligation for Administrations to publish the source code of their algorithms, a criterion that is part of the defence of the virtues of open source. However, there is still a lack of consensus around this question.

Open Source Code – Proprietary Source Code

Open source

- Open source software refers to software that allows its users **unrestricted access to its source code**.
- Open source licenses also allow unrestricted distribution of code for all purposes, and are often provided **for free**, although there are also viable business models in relation to the development and implementation of open source software.
- The purpose of sharing code is to allow the community to test the code for **errors and possible security risks** so that it can be further improved. This greatly improves the quality of the software. In relation to the ethics of algorithms, it is an option that may eventually allow the identification of biases and other unwanted effects.
- However**, not everyone has the knowledge needed to test code**, and there may be people within this expert minority who take advantage of the situation and exploit their knowledge of open source systems to their own benefit.

Proprietary Code

- Proprietary code **is owned by the creator** and is their legal property, and offers a restricted view of their technical operation.
- **The code is hidden** and the software itself often has to be purchased as well, which prevents the software infrastructure from being exposed to cybercriminals.
- However, this does not make the software completely immune to security risks, which **cannot be verified because it is not possible to see the code.** In addition, the user would have to blindly trust their software provider.
- There are very strict conditions regarding the use of this type of software, and its **distribution** is also often **prohibited**.
- For this reason, it is common for software developers to produce proprietary code, as they consider that it helps them make a financial profit from the code and thus recover their investment in research and development. Therefore, **many advanced software features are only found in proprietary code**.

On the other hand, **the fact that the algorithm code is published does not necessarily mean that it can be understood.** The parameters of a machine learning algorithm are adjusted by the machine, and humans do not have the cognitive capacity to find meaning in them, not even the people who have developed the system. They are therefore what are known as **“black boxes”**.

2. Tools for algorithm harm prevention

Transparency: explicability

Interpretability

- A system is interpretable if, simply by observing the elements that make it up, an expert is able to understand its operation and predict what output it will give for certain inputs.
- Today's complex Artificial Intelligence models are able to perform very complex tasks with highly accurate results, but humans **are not capable of interpreting them**. These models look at multiple aspects of the inputs and process them through complex operations with multiple parameters, making it impossible to directly appreciate which variables contribute most to the algorithm making one or another decision.
- This lack of interpretability affects the trust of institutions and users in AI.

Explicability

- When it is not possible to directly interpret an algorithm, a more modest alternative is to give an approximate explanation of the behaviour of the complex system by means of a simpler, interpretable system, i.e. one which is understandable and predictable for a human.
- Thus, an attempt is made to approximate the behaviour of a complex algorithm as a weighted sum of factors.
- To calculate the weight of each factor, explicability techniques "play" with the algorithm by observing how its output varies with variations in the input. In order to perform such an analysis, the supplier of the analysed algorithm needs to provide certain deliverables.

Explicability techniques for deciphering black boxes

Local explicability techniques, such as LIME, explain why the algorithm has made a particular decision. For example, if an algorithm considered an email to the SAU service to be about a user registration request for an application, LIME would pinpoint the words in the email that caused the algorithm to classify it that way.

Global explicability techniques, such as SHAP, explain what factors influence all the algorithm's decisions. In the example of the SAU email classification algorithm, SHAP would show which words lead an email to be classified as an incident, request or query.

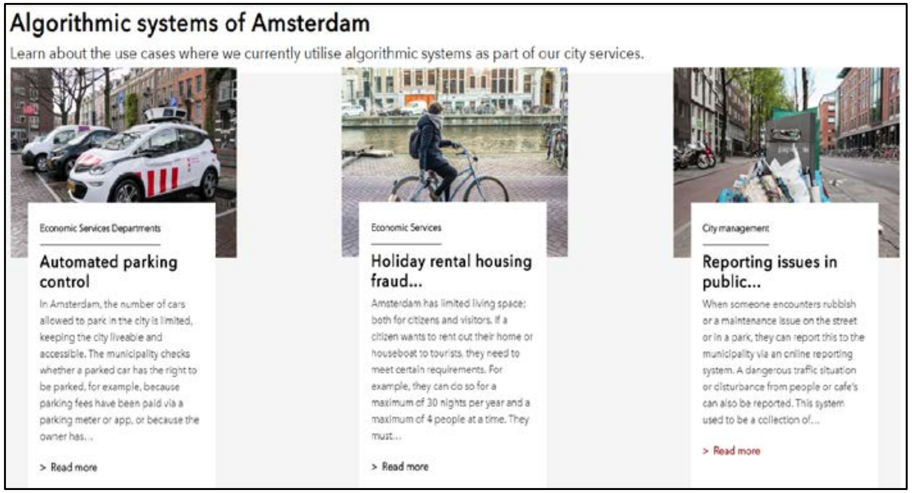
2. Tools for algorithm harm prevention

Transparency: traceability

An alternative approach to the principle of transparency is the one taken by the town councils of **Amsterdam and Helsinki,** which have recently (September 2020) become the first cities to **publish open registries of the AI algorithms** they use.

On their specific portals, both cities provide an overview of each AI system, identifying the problem it was intended to solve and why a particular algorithm was selected for the task. Details are also provided on the training data, the data used and their operational logic, as well as the risks identified and the precautions taken.

The key idea is that however opaque AI systems may be, that does not make the institutions promoting them any less accountable, and said institutions can do something very important to ensure their transparency: they can **trace all the decisions and actions that have been taken with respect to each algorithm**, thus becoming **fully accountable** for their actions and sharing their knowledge with the public.



Algorithmic systems of Amsterdam

Learn about the use cases where we currently utilise algorithmic systems as part of our city services.

Economic Services Departments

Automated parking control

In Amsterdam, the number of cars allowed to park in the city is limited, keeping the city liveable and accessible. The municipality checks whether a parked car has the right to be parked, for example, because parking fees have been paid via a parking meter or app, or because the owner has...

> Read more

Economic Services

Holiday rental housing fraud...

Amsterdam has limited living space: both for citizens and visitors. If a citizen wants to rent out their home or houseboat to tourists, they need to meet certain requirements. For example, they can do so for a maximum of 30 nights per year and a maximum of 4 people at a time. They must...

> Read more

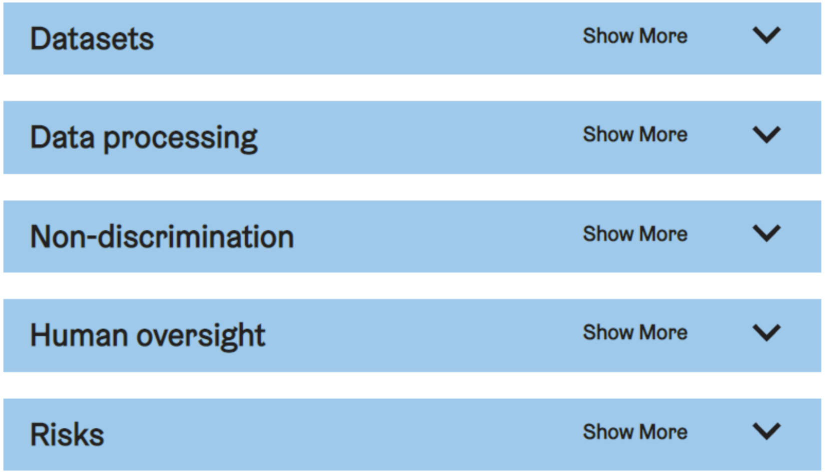
City management

Reporting issues in public...

When someone encounters rubbish or a maintenance issue on the street or in a park, they can report this to the municipality via an online reporting system. A dangerous traffic situation or disturbance from people or cafe's can also be reported. This system used to be a collection of...

> Read more

Amsterdam AI Algorithm Portal: https://algorismeregister.amsterdam.nl/en/ai-register/



Details of an algorithm on the Helsinki portal: https://ai.hel.fi/en/ai-register/

"Algorithms play an increasingly important role in our lives. Together with the city of Helsinki, we are on a mission to create as much understanding about algorithms as possible and be transparent about how we – as cities – use them." *Touria Meliani, Deputy Mayor of Amsterdam.*

3. Government algorithm harm prevention frameworks

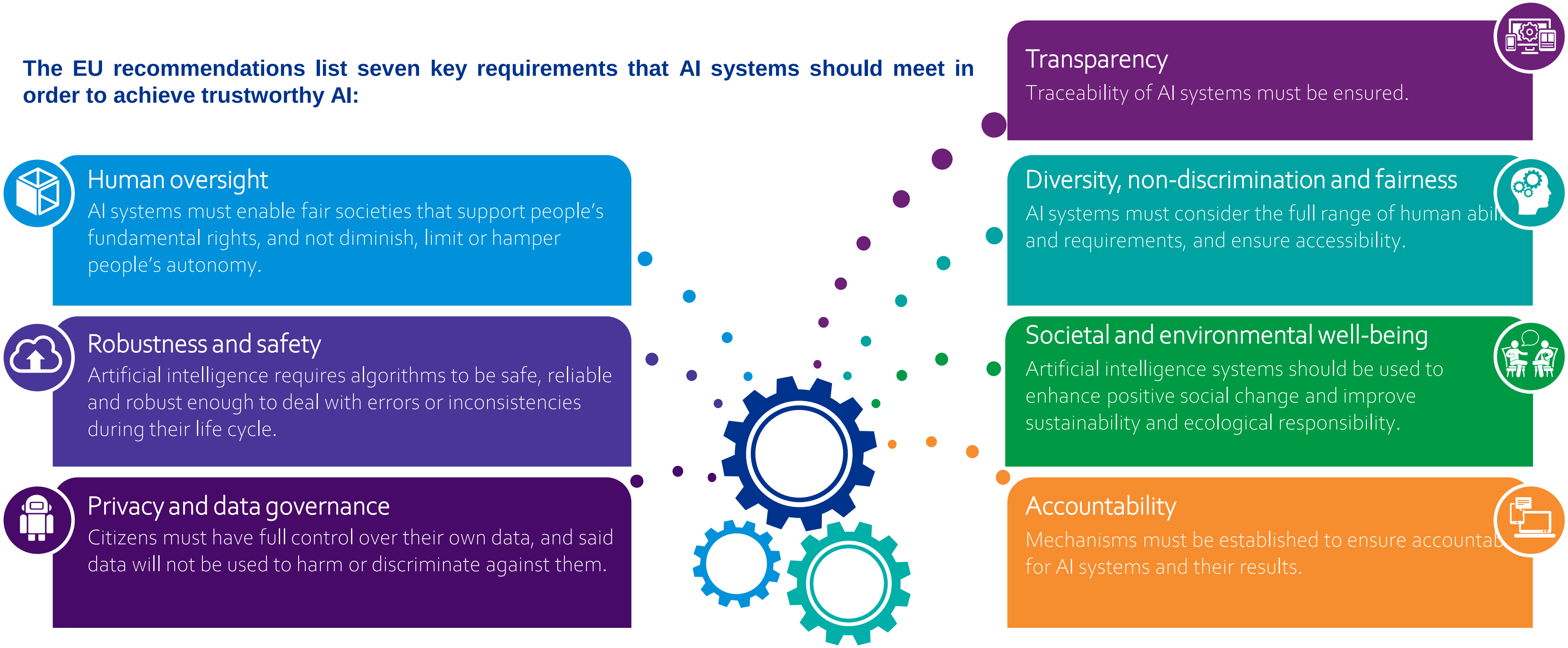
Each of the instruments that have been shown make it possible to detect and, eventually, help to correct some of the risks associated with algorithms, but none of them can be used by itself to separate all the desirable algorithms from all the undesirable ones.

For this reason, every organization that implements algorithms should establish a framework, i.e. a solid system of instruments and mechanisms to prevent harm from algorithms. Below, we give several examples of what these frameworks might look like, especially for the case of governments that seek to ensure the ethics of the algorithms implemented by their public services.

International recommendations for governments: European Commission

In 2019, the European Commission published its non-binding **recommendations for trustworthy AI**. The purpose of these recommendations is for Member States to implement frameworks to ensure the ethical development of artificial intelligence, both in the public and private sectors, and identify the content that these frameworks should include.

The EU recommendations list seven key requirements that AI systems should meet in order to achieve trustworthy AI:



Human oversight
AI systems must enable fair societies that support people's fundamental rights, and not diminish, limit or hamper people's autonomy.

Robustness and safety
Artificial intelligence requires algorithms to be safe, reliable and robust enough to deal with errors or inconsistencies during their life cycle.

Privacy and data governance
Citizens must have full control over their own data, and said data will not be used to harm or discriminate against them.

Transparency
Traceability of AI systems must be ensured.

Diversity, non-discrimination and fairness
AI systems must consider the full range of human abil and requirements, and ensure accessibility.

Societal and environmental well-being
Artificial intelligence systems should be used to enhance positive social change and improve sustainability and ecological responsibility.

Accountability
Mechanisms must be established to ensure accountab for AI systems and their results.

3. Government algorithm harm prevention frameworks

International recommendations for governments: Toronto Declaration

In May 2018, Amnesty International, Access Now and other partner organizations presented the **Toronto Declaration on the protection of the right to equality and non-discrimination in machine learning systems**. The Declaration is a landmark document that seeks to apply existing international human rights standards to the development and use of artificial intelligence systems.

The document sets out the following measures that should be taken by states to mitigate and reduce the harm of discrimination from machine learning in public sector systems. **The measures can be structured into three blocks:**

1 IDENTIFY RISKS

Any state deploying machine learning technologies must thoroughly investigate systems for discrimination and other rights risks prior to development or acquisition. States should:

- **Carry out periodic impact assessments** during all stages of the project to identify potential sources of discriminatory or other rights-harming outcomes.
- **Take appropriate measures** to mitigate risks identified through:
  - Impact assessments.
  - Conducting pre-release trials.
  - Ensure that potentially affected groups and field experts are included as actors with decision-making power.
  - **Independent expert review**, where appropriate.
- **Disclosing known limitations** of the system in question, for example, noting measures of confidence, known failure scenarios and appropriate limitations of use.

2 ENSURE TRANSPARENCY AND ACCOUNTABILITY

States must ensure and require accountability and maximum possible transparency around public sector use of machine learning systems. **This must include explicability and intelligibility in the use of these technologies** so that the impact on affected individuals and groups can be effectively scrutinised by independent entities. States should:

- Disclose where machine learning systems are used in the public sphere, **providing information that explains in clear and accessible terms how decision-making processes are reached** and document the actions taken.
- Enable independent analysis and oversight through **by using systems that are auditable.**
- **Avoid the acquisition and use of 'black box systems'** that do not provide the required information or do not allow the established explicability techniques.

3 ENFORCE OVERSIGHT

States must take steps to ensure public officials are aware of and sensitive to the risks of discrimination and other rights harms in machine learning systems and **ensure oversight**. States should:

- Include conditions in the **tender specifications** so that those involved in the design, implementation and review of machine learning include the necessary steps in the supply or development and maintenance process.
- Ensure that public bodies carry out **training in human rights and data analysis for officials involved** in the procurement, development, use and review of machine learning tools.
- Create **mechanisms for independent oversight**, including by judicial authorities when necessary.
- Ensure that machine learning-supported decisions **meet international accepted standards**.

3. Government algorithm harm prevention frameworks

Automated decision-making and impact assessment in Canada

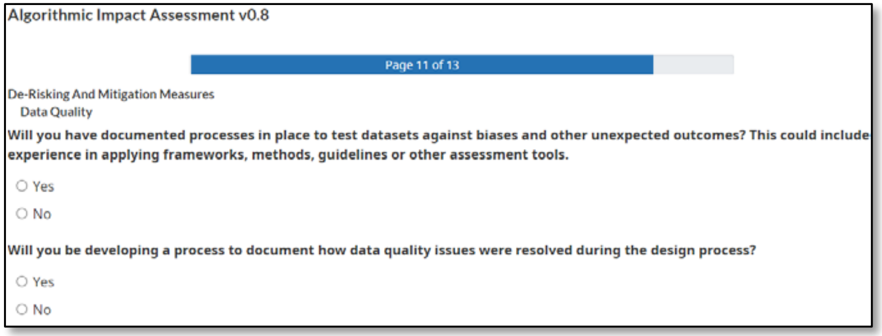
Canada's Directive on Automated Decision-Making – Algorithm Impact Assessment

This directive regulates any algorithm used to make or recommend an administrative decision.

It divides decisions into **4 levels according to their impact** (little to none, moderate, high and very high) on the rights, health or economic interests of people, organizations and communities, or the sustainability of an ecosystem. For each level, the directive imposes requirements such as:

- Peer review; the directive specifies how many experts and from what areas.
- Publishing a news item about how the system works.
- Human intervention in the decision-making process.
- Explicability of decisions.
- Initial testing and periodic review.
- Monitoring of the system's decisions.
- System documentation and certified user training.
- Contingency plan and emergency backup systems.
- Approval by an administrative body; the directive specifies which one.

Impact assessment is provided using an Algorithmic Impact Assessment (**AIA)** tool, a questionnaire to be completed by departments and agencies that wish to implement an algorithm **with around 60 questions related to the processes, data and decisions made by the AI system in question**. The tool extracts results that demonstrate the system's level of impact and also specifies the requirements for the system in question according to the applicable legislation. The information from the AIA is only stored locally on the user's computer and the Canadian Government does not have access to the information entered in the tool.



Algorithmic Impact Assessment v0.8

De-Risking And Mitigation Measures

Data Quality

Will you have documented processes in place to test datasets against biases and other unexpected outcomes? This could include experience in applying frameworks, methods, guidelines or other assessment tools.

○ Yes

○ No

Will you be developing a process to document how data quality issues were resolved during the design process?

○ Yes

○ No

3. Government algorithm harm prevention frameworks

UK Data Ethics Framework

1 Approach

The United Kingdom has addressed the issue by publishing a series of recommendations for the appropriate and responsible use of data in the public sector. The scope includes algorithms, which operate on data, but also other areas such as statistics and data science. The approach chosen is voluntary guidance to help the department or agency throughout the entire data use project.

The guide is based on the principles of transparency, accountability and fairness, and refers to the principles of the Toronto Declaration, but has a very practical approach based on specific actions.

2 Steps

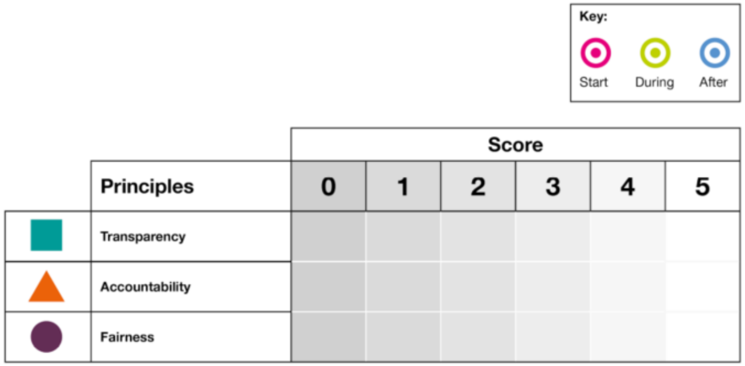
The Framework proposes a 5-step method, with each step dealing with questions such as:

1. Understand the project and its context
    - Which individuals and groups will benefit or be harmed by the project?
    - What unintended consequences might it have, and how can they be avoided?
    - How does it affect human rights?
2. Involve diverse expertise
    - Build a diverse expert team
    - Involve external stakeholders in the project and its governance
    - Publish the consultations made
3. Comply with the law (on data protection, equality, etc.).
4. Review the quality and limitations of the data
    - Data sources
    - Biases
5. Assess and consider implications for public policy
    - Did it work as expected?
    - Have user needs changed?
    - ...

3 Self-assessment

Each of the 5 steps in the guidelines includes a series of aspects to consider. These are available in checklist format, in a downloadable tool that provides a score between 0 and 5 for the 3 principles of transparency, accountability and fairness.

If a project scores 3 or less in any area, the guidelines recommend seeking advice from the data ethics officer of the department or agency responsible.



Key: ◎ Start ◎ During ◎ After

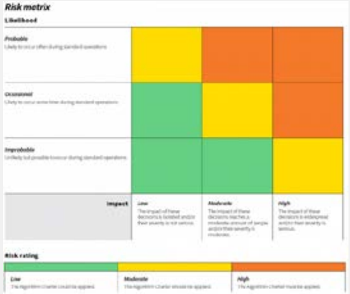
		Score					
	Principles	0	1	2	3	4	5
■	Transparency						
▲	Accountability						
●	Fairness						

3. Government algorithm harm prevention frameworks

Other high-level state regulations

New Zealand has published the world's first "Algorithm Charter". It is a set of principles to which the different ministries and other public bodies can adhere in relation to the implementation of algorithms in public services.

The strategy taken in the charter is to focus on decisions that involve more risk, i.e. those that are more likely to produce a high impact.


Risk matrix
Likelihood
Probable
Likely to occur often during standard operations
Occasional
Likely to occur some time during standard operations
Improbable
Unlikely but possible to occur during standard operations
Impact
Low
The impact of these decisions is isolated and/or their severity is not serious.
Moderate
The impact of these decisions reaches a moderate amount of people and/or their severity is moderate.
High
The impact of these decisions is widespread and/or their severity is serious.
Risk rating
Low
The Algorithm Charter could be applied.
Moderate
The Algorithm Charter should be applied.
High
The Algorithm Charter must be applied.

The commitments of the signatory bodies include:

- **Transparency**, through the publication of documentation on the algorithm and its data.
- **Participation** of the people and communities concerned, through public consultations and, in particular, the inclusion of the Maori perspective.
- Identification, understanding and management of **data** limitations and biases.
- Protection of privacy, ethics and human rights through **periodic peer reviews**.
- **Human supervision**, either in each individual decision, or in the form of a contact channel to receive public inquiries about the algorithms or appeals against their decisions.

The United States is working on an algorithm law called the Algorithmic Accountability Act.

The law would apply to subjects:

- With an annual turnover of more than $**50M**
- Or with **information on more than 1M people or devices**

In **automatic decision-making systems considered high-risk** due to:

- Using new technology
- The scope of the service
- Profiling to predict people's behaviour
- The use of protected personal information such as race or political opinions
- Or because they monitor public spaces.

The risks considered are diverse, so the law would prescribe:

- A data protection impact assessment
- An **impact assessment of the automated decision system**, which would include:
    - Description of the system, its purpose and its data
    - Assessing the system's costs and benefits, considering factors such as data minimization practices or consumer access to data
    - Assessment of risks
    - Measures to avoid them