



Algorithm Ethics Framework

Ministry of Digital Policy and Public Administration

Generalitat de Catalunya

Dossier for the ApLab session of 19 January 2021





Contents

1. Introduction

2. Building the ethics of algorithms

3. *Basement: foundations*

B1. Awareness and training

B2. Caution

B3. Proportionality

4. *Floors: approval of algorithms*

F1. Socio-technical analysis

F2. Data protection assessment

F3. Specifications

F4. Analysis of bias and global explicability

F5. Approval by resolution

6. *Attic floors: guarantees for algorithms already in use*

A1. Transparency: algorithm data sheet

A2. Human oversight

A3. Questioning decisions



1. Introduction

2. ApLab work flow

3. Building the ethics of algorithms

1. Introduction

1 Proactive and personalized services can help governments better serve citizens, but the use of algorithms involves several risks. In the previous ApLab, different elements were analysed and prioritized to prevent risks related to algorithms in the Public Administration.

2 These methods provide the Generalitat with an **algorithm ethics framework**: a structured set of criteria and practices to guarantee the responsible implementation of algorithms in public services. These elements must cover the entire algorithm life cycle:

- ✓ General criteria
- ✓ Algorithm approval procedures
- ✓ Guarantee mechanisms for algorithms that are already implemented

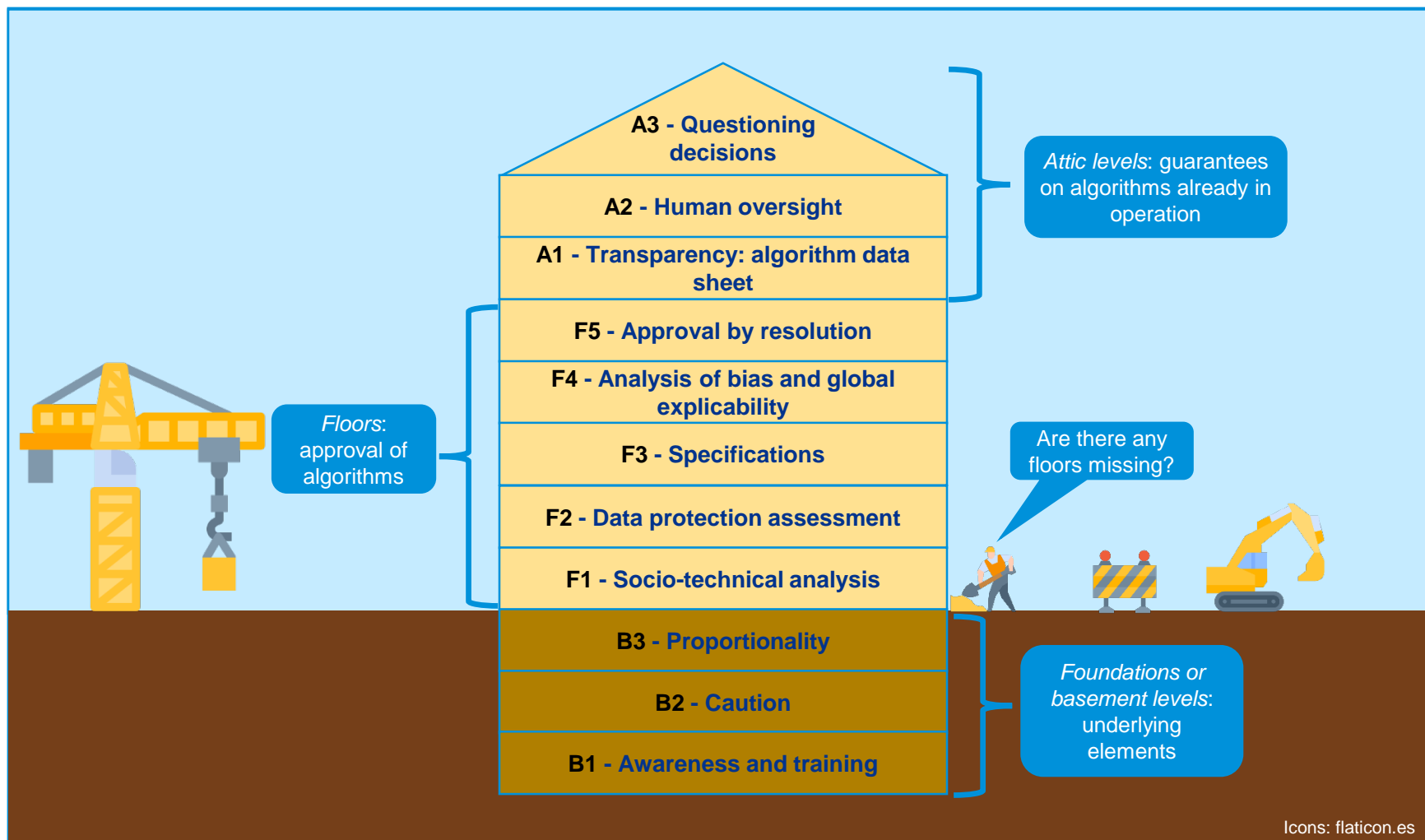
3 The purpose of this dossier is to help define this framework. For each of the 11 prioritized elements, the following questions are answered:

- ✓ What does the element consist of?
- ✓ What questions should be defined?
- ✓ International references



2. Building the ethics of algorithms

We will represent the elements that intervene in algorithm ethics, under the metaphor of a building with three basement levels and eight floors:





3. Basement levels: foundations

B1. Awareness and training



B2. Caution

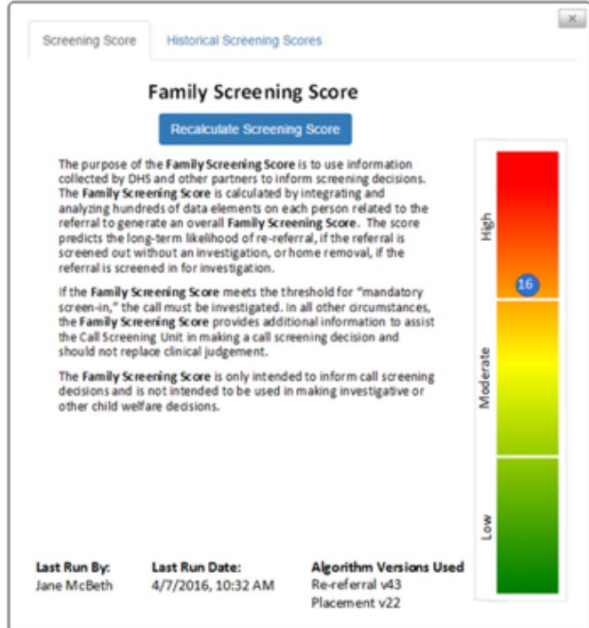
B3. Proportionality

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

3. Basement levels - foundations

B1. Awareness and training

 What does this consist of?	 Questions to address during the session
<p>Awareness and training on algorithms and their risks. It would be necessary to train and sensitize all public workers who participate in its design, supply, implementation and operation, so that they identify the risks and learn how to mitigate them.</p>	<ul style="list-style-type: none"> • Which groups should receive general training and sensitization? • In what format should it be done? • Where should the greatest emphasis be placed?

References	
<p>1. The Allegheny County AI User Training Programme - link</p> <p>The Allegheny Family Screening Tool is an algorithm that helps Social Services identify children who are potential victims of abuse, in order to take preventive measures that might include the suspension of parental rights. It suffers from a data collection bias, since it is easier for wealthy families to hide abuse, as injuries can be treated through private medical insurance.</p> <p>The staff working with this algorithm have received specific training which has allowed them to identify such possible cases where data has been hidden. This training has allowed them to correct the bias produced by the algorithm.</p>	 <p>The screenshot shows the 'Family Screening Score' interface. At the top, there are tabs for 'Screening Score' and 'Historical Screening Scores'. Below the tabs is the title 'Family Screening Score' and a 'Recalculate Screening Score' button. The main content area contains text explaining the purpose of the score and a vertical color scale from Low (green) to High (red). The score 16 is displayed on the scale. At the bottom, there is a summary of the last run: 'Last Run By: Jane McBeth', 'Last Run Date: 4/7/2016, 10:32 AM', and 'Algorithm Versions Used: Re-referral v43, Placement v22'.</p>

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

3. Basement levels - foundations

B1. Awareness and training

References:

2. Introductory AI course for governments

The training programme of the UK's GDS Academy (school for public employees of the British Government Digital Service), open to all the country's public servants, includes a 3-hour course on [AI for the public sector](#).

The Swedish Association of Municipalities and Regions offers an [AI training activity for public sector leaders](#), for the elected officials and managerial staff of public administrations. It consists of an individual study part and a one-day online workshop.



Government
Digital Service

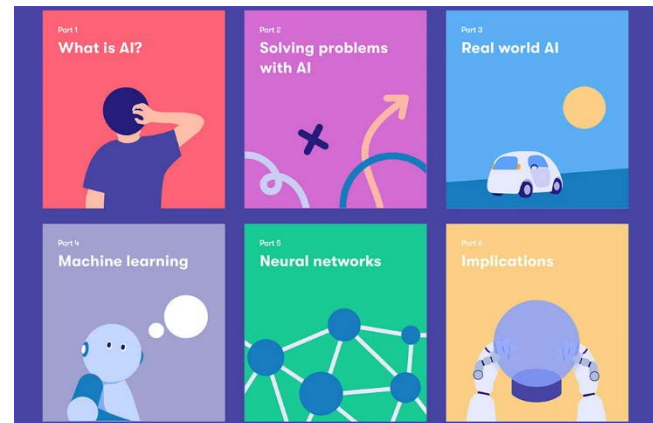


Sveriges
Kommuner
och Regioner

3. The Finnish massive AI training strategy – link

The plan to position Finland as a leading country in artificial intelligence includes the objective of **making a large part of the Finnish population literate in artificial intelligence**, to ensure citizens have a basic understanding of the AI applications around them and to adapt the population to the needs of the labour market.

The core element of this strategy is the massive online course "[Elements of AI](#)", which in just one year was taken by 3% of the Finnish population and which has reached 550,000 people around the world. This basic course can be followed up with additional courses, including a specific one on [AI ethics](#).



A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

3. Basement levels - foundations

B2. Caution

What does this basement level consist of?

The Public Administration has a lot of room for digital transformation through established technologies, which overcome the challenge of interoperability by organizing data to make decisions based on simple rules. In sensitive areas, safety must be prioritized over innovation with immature technologies that may produce unacceptable risks. This precautionary principle is applied as a **partial moratorium** on certain technologies in certain areas.

Questions to address during the session

- Which technologies should be restricted?
- In which areas should immature technologies be avoided?

References

1. Facial recognition banned in French high schools - [link](#)

Concerns have been raised around the use of facial recognition technology for reasons such as the capacity for totalitarian State control and racial bias. Companies such as IBM have stopped developing this technology for these reasons.

A French court banned its use for the purpose of controlling high school attendance, considering that the power relationship between school and student prevented free consent, which, according to Article 22 GDPR, is one of the exceptions that make it possible to perform automated decision-making and profiling.

A [European Parliament report](#) proposes a **moratorium on the use of facial recognition systems** until these systems can be understood to respect fundamental rights and comply with the applicable regulations, without leading to discriminatory results, and there is trust in the necessity and proportionality of these systems.




A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

3. Basement levels - foundations

B2. Caution

References

2. Proposed moratorium on uses of AI in the US judicial system – [link](#)

“Partnership on AI”  a consortium of the main tech companies, with the participation of representatives from universities and civil society, to establish best practices for AI systems – has published a report on the currently available crime risk assessment algorithms.

According to the report, these algorithms are currently unreliable, so **they should not be used to make decisions regarding arrests**. They do, however, recommend that they be used to speed up decisions on the release of prisoners in the context of the United States, which has a much higher incarceration rate than all other developed countries. However, when used, algorithms will need to report their **margin of error**, and users will have to receive **training to avoid biases**.



3. European criteria for limiting high-risk AI applications – [link](#)

In February 2020, the European Commission published its “White Paper On Artificial Intelligence - A European approach to excellence and trusts”. The AI applications considered to be high-risk in this strategic document are:

- Those in sectors where there is risk: judicial system, social services, employment services, health, etc.
- And where the way AI is used involves a significant risk: death, injury or significant material or immaterial damage.



Some applications, such as intrusive surveillance or personnel selection processes, will be considered high-risk regardless of the sectors in which they are applied.




A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

3. Basement levels - foundations

B3. Proportionality

 What does this consist of?	 Questions to address during the session
<p>The controls to apply in the implementation of algorithms will have to be graduated in proportion with the risk involved in each case, in terms of the data used and the possible effects.</p> <p>However, in the case of modifications to systems already in operation, which frequently undergo technical modifications, the controls will have to be more minor than in the case of new algorithms.</p>	<ul style="list-style-type: none"> • Which objective criteria determine which filters to apply? • Which criteria will be subject to assessment? • Who will perform this assessment?

References

<p>1. Bill of the US “Algorithmic Accountability Act” – link</p> <p>The US Senate is processing a bill that would impose audits on the most relevant automated decision-making systems. The rule would impose assessments on:</p> <ul style="list-style-type: none"> • Companies with an annual turnover of >\$50M • Or which hold information from >1M users • High-risk automated decision-making systems <p>New technologies defined as high-risk are:</p> <ul style="list-style-type: none"> • Systems that, due to the newness of their technology, context or purpose, involve a risk to privacy, or that produce unfair decisions that impact consumers. • Those that make decisions based on consumer profiling • Those involving sensitive data • Those that monitor public spaces 	
---	--


3. Basement levels - foundations

S3. Proportionality

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

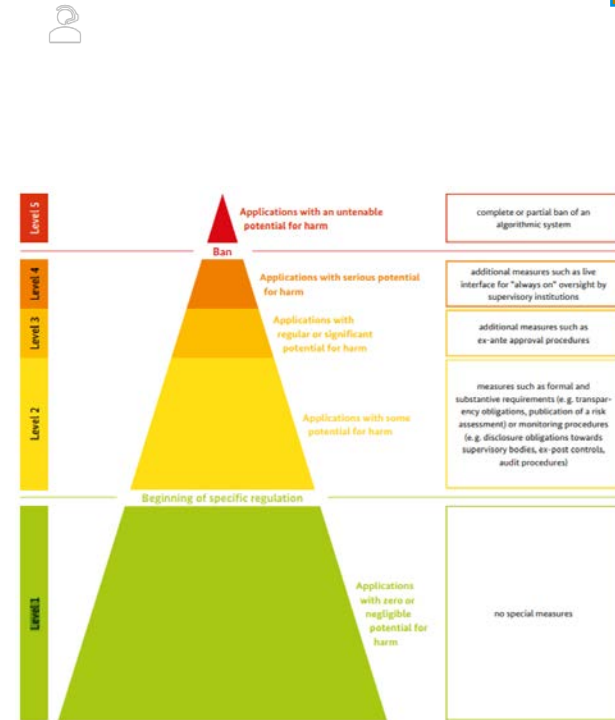
References

2. German Data Ethics Commission criticality pyramid – [link](#)

This government  commission has proposed to **classify** algorithm-based systems (in both the public and private sectors) **into 5 levels according to their potential to cause harm**, and for each level it establishes the type of measures to be applied:

1. Zero or negligible risk (e.g.: vending machine) – no special measures required.
2. Some risk (e.g.: intelligent calculation of mobility routes) – risk assessment, transparency, and ex-post controls in the event of suspected inappropriate operation.
3. Significant risk (e.g.: establishment of personalized prices) – ex-ante approval procedure, accompanied by a periodic review.
4. Serious risk (e.g.: banking systems that assess people applying for credits) – “always on” oversight by the institutions.
5. Unsustainable risk (e.g.: weapons that determine their targets autonomously) – complete or partial ban.

The risk includes the **probability of harm and its severity**, and should not be assessed for the algorithm in isolation but **for the socio-technical system** as a whole, which includes all the people involved, from development to the ordinary operation and assessment of the system.





4. Floors: approval of algorithms

F1. Socio-technical analysis

F2. Data protection assessment

F3. Specifications

F4. Analysis of bias and global explicability

F5. Approval by resolution

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

4. Floors – steps for the approval of algorithms

F1. Socio-technical analysis

What does this step consist of?

Socio-technical analysis, by means of a multidisciplinary team that analyses how a complex social problem has been reduced to data processing, and identifies the risks, the groups to be protected, the problems linked to the data and the possible mitigation strategies. Universities have capabilities that can facilitate this task.

Questions to address during the session

- What content needs to be analysed?
- What professional profiles do we need for the implementation of the Generalitat's algorithm strategy?
- Where do we get these professionals from?

References

1. **Eticas Research & Consulting algorithm audit methodology – [link](#)**

The Eticas consultancy has developed a methodology to audit the ethics of algorithms, which includes these 3 steps:

1. Socio-technical analysis, **to understand** how a company or government has **reduced a complex social issue to data processing**. Often, organizations do not use the data they need, but the data they have, and this conditions the ethical impact of the algorithm.
2. Technical analysis to find out **how the algorithm works** and identify the **vulnerable groups** and the impact on them.
3. Analysis of the **interaction between the result of the algorithm and the human input**. In Europe, the General Data Protection Regulation requires that the final decision be taken by a person (human in the loop), but algorithm bias and human bias can form a negative combination.




4. Floors – steps for the approval of algorithms

F1. Socio-technical analysis

References

2. Analysis of a tourist apartment fraud prevention algorithm – [link](#)

Amsterdam City Council  has piloted an algorithm to help identify violations of tourist apartment regulations. The system processes the complaints received to prioritize the limited resources of the inspection team, based on data regarding residents, apartments and the records of fraud cases throughout the city.

All the data sets were analysed to exclude all attributes that could lead to discrimination (e.g. nationality). However, it has been detected that the algorithm can still distinguish between certain social groups, by means of other attributes such as post code or the number of members in the family unit. For this reason, this bias will be researched further during the pilot.



3. Digital service design standard in the UK – [link](#)

The Service Standard is the set of principles that set out how the UK Government's digital services should be designed. These principles include:

- Understanding users and their needs
- Ensuring that no group of users is excluded

The standard specifies that services must be designed by small multidisciplinary teams, and includes descriptions of the different roles. One of the roles is as follows:

- User researcher: ability to understand the social and technological context, the diversity of users and the problem to be solved, and user research methodology and data analysis

The UK Government's Centre for Data Ethics and Innovation (CDEI), in its [report on biases](#), mentions the need to form socially diverse teams.



Government Digital Service


A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

4. Floors – steps for the approval of algorithms

F2. Data protection assessment

 What does this step consist of?	 Questions to address during the session
<p>The European legislation on data protection establishes two risk analysis instruments for systems regarding personal data and the design of mitigating measures: the assessment of the level of risk and, in case of high risk, the data protection impact assessment.</p>	<ul style="list-style-type: none"> Balancing data minimization and bias mitigation

References

<p>1. The EU General Data Protection Regulation (RGPD) – link</p> <p>It is the most prestigious data protection regulation globally. It defines the obligations of data controllers and processors to guarantee the rights of users, taking a proportional risk-based approach.</p> <p>The Spanish Data Protection Agency has published a Guide to compliance with the GDPR for processing that incorporates artificial intelligence. AI systems pose specific challenges:</p> <ul style="list-style-type: none"> The data controller must ensure that a mature technology is implemented that is precise, accurate and predictable, and which allows the legal requirements of responsibility and transparency to be met. If personal data are used, the training of the model, its validation and its operation are different types of processing, with different legitimate purposes. The UK Government report on algorithmic biases points out that the principle of data minimization clashes with that of bias minimization: if we do not save attributes such as nationality, we will not be able to know whether the algorithm discriminates against people defined by this attribute. 	
--	--

4. Floors – steps for the approval of algorithms

F2. Data protection assessment

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

References

2. Automated decision-making

Article 22 of the RGPD establishes that everyone will have the right not to be the subject of a decision based solely on automated processing, which produces legal effects or significantly affects them. This means that consent is therefore required.

For consent to be valid, it must be free and informed, i.e. the person must have alternatives and information about the logic applied. This requirement does not refer to the source code but to data with relevance to the person's decision, such as:

- Data used and storage period
- Importance that each piece of data has on the decision
- Accuracy metrics, audits and certifications

An alternative, so that this consent is not required, is for the final decision to be made not by the machine but by an employee, what is known as “human in the loop”. In this case it must be guaranteed that the employee's role will not be limited to always approving the machine's proposal, but that they will also be trained to have a well-founded opinion.



3. Profiling

The same precept applies to systems that perform profiling, i.e. processing that seeks to infer information about a person by analysing or predicting personal characteristics.




If profiling is performed, the system must undergo a Data Protection Impact Assessment, which is a formal risk analysis that must set out the risks and measures taken, and if there are still significant risks after these measures have been applied, the authorities should be consulted.



A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

4. Floors – steps for the approval of algorithms

F3. Specifications

 What does this step consist of?	 Questions to address during the session
<p>Request for proposals based on technical specifications that incorporate criteria related to fairness and the related guarantee mechanisms.</p> <p>It will be necessary to update the technical criteria based on the current state of technology and society, as customs and requirements are not</p>	<ul style="list-style-type: none"> • Who has the ability to draft them? • What clauses might they contain? • Who can assess the proposals? • Are all algorithms procured via contract?
References	
<p>1. Explainable AI in the Community of Madrid – link</p> <p>In 2018, the Community of Madrid opened a call for tenders to supply information systems to carry out a digital transformation based on data and artificial intelligence. In the design, development, implementation and maintenance contract that was tendered the bidders were asked to propose a XAI (eXplainable Artificial Intelligence) strategy which included the procedures, methodology, organization and technological measures to overcome bias in AI models.</p> <p>The specific administrative specifications established a maximum of 7 points for the evaluation of this proposal, within the technical proposal evaluation section, corresponding to a qualitative criterion based on a value judgement.</p>	

4. Floors – steps for the approval of algorithms

F3. Specifications

References

2. AI Procurement in a Box – [link](#)

This summer, the World Economic Forum published a guide to help governments in the procurement of artificial intelligence technologies, according to innovation, efficiency and ethics criteria.

The toolkit contains:

- A list of questions to help draft requests for proposals
- A guide to determine which sections of the list need to be emphasized for different types of specifications
- A list of guidelines for successfully carrying out an AI procurement process.
- International examples of success stories

The guidelines recommend that the contracting process be **preceded by a data protection risk and impact assessment**, to then be able to **ask potential suppliers how they would resolve the risks** detected. It is recommended that an **iterative approach** be taken that makes it possible to get to know and master the technology, and that the contracted solutions be accompanied by the transfer of knowledge to the technological managers and functional users of the organization where it is implemented.

The objectives for the specifications include privacy, the avoidance of bias, explicability and the possibility of system drift, as the most ethically relevant aspects.

6 Concept drift: Ensuring the system does not drift from its intended purpose

Sample specification

6.6 Explain how you will ensure the AI system or service does not drift from its intended purpose or outcome.



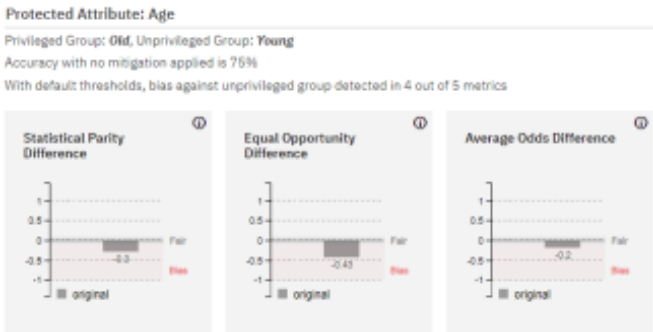
Key considerations to look out for in the answers

56. As algorithms are learning continuously after they are developed it is possible for them to drift from the original concept and deliver different results. Providers can be assessed on their approach to the following:
- What is the expected performance on unseen data or data with different distributions?
 - Does the system make updates to its behaviour based on newly ingested data?
 - Is the new data uploaded by users? Is it generated by an automated process? Are the patterns in the data largely static or do they change over time?
 - Are there any performance guarantees/bounds?
 - Does the service have an automatic feedback/retraining loop or is there a human in the loop?
 - How is the service tested and monitored for model or performance drift over time?
 - Is the supplier providing performance drift monitoring KPIs that prompt retraining if there are any unexpected changes?
 - How can the service be checked for correct, expected output when new data is added?
 - Does the service allow for checking for differences between training and usage data?
 - Does it deploy mechanisms to alert the user of the difference?
 - Do you test the service periodically?
 - Does the testing include bias or fairness related aspects?
 - How has the value of the tested metrics evolved over time?

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
P4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

4. Floors – steps for the approval of algorithms

P4. Analysis of bias and global explicability

 What does this step consist of?	 Questions to address during the session
<p>Once a complex and opaque algorithm has been developed, its operation should ideally be characterized by means of an analysis of biases in the result and global explicability techniques.</p>	<ul style="list-style-type: none"> Who should perform the analysis? What is needed in order to perform the analysis?
References	
<p>1. IBM’s “AI-Fairness-360” tool – link</p> <p>This is an open-source tool developed by IBM for free use, to detect bias and discrimination in the application of AI algorithms.</p> <p>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.</p> <p>Biases in the input data can be corrected by obtaining fairer real data, or by means workarounds such as the generation of artificial data or the reweighting of the various groups. The application makes it possible to apply various bias mitigation techniques and see what improvements they bring in the five previously measured fairness metrics.</p> <p>The objective of toolkits like AI Fairness 360° is not to distinguish between good and bad algorithms but to help the institutions that create them to make them fairer. It is therefore intended for use during development rather than as a final test of approval by the institution that acquires the algorithm, although it can also be used for this purpose.</p>	

4. Floors – steps for the approval of algorithms

F4. Analysis of bias and global explicability

References

2. UK Government guidance on mitigating bias – [link](#)

The UK Government's Centre for Data Ethics and Innovation (CDEI) has published guidance on identifying and mitigating bias in the use of artificial intelligence in the public sector.

Fairness is not an attribute of the algorithm but of the system it is part of, and it is a complex concept that goes beyond the absence of bias. The guide illustrates different understandings of fairness, which can be contradictory.

A system directly discriminates when it makes a decision according to a protected variable; this can be avoided by not collecting it – this is what is called *fairness through unawareness*. But it is often not effective in avoiding bias, since discrimination can also occur indirectly, privileging some groups over others through unprotected variables.

The guide recommends identifying biases and correcting them, rather than hoping to avoid them through unawareness. To do this, it presents a range of statistical analysis tools, as well as correction strategies that can be applied to the input data, the operation of the algorithm or its outputs.

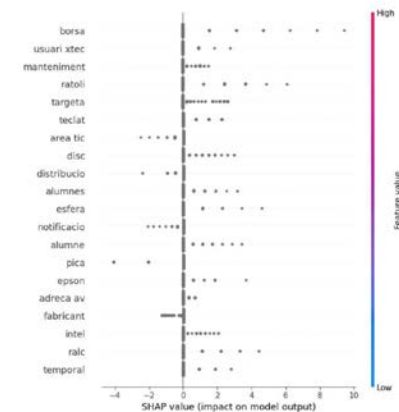


3. Global explicability

Global explicability techniques, such as SHAP, show which factors influence all the algorithm's decisions, assigning a weight to each factor. This allows us to see whether the algorithm takes into account only factors that seem reasonable, or if it makes decisions based on variables that privilege some groups over others.

It must be said, however, that this calculation of weights is an approximation. SHAP works well for weighting independent criteria, but not so well when the decision follows more complex rules.

The Generalitat is developing a tool that would allow local explicability techniques such as LIME and global ones such as SHAP to be applied to artificial intelligence algorithms.



A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

4. Floors – steps for the approval of algorithms

F5. Approval by resolution

What does this step consist of?

Approval of the algorithm, in the case of automated administrative actions, by means of a **resolution** which specifies the assurance mechanisms available to the people affected by the algorithm.

In the case of algorithms with a greater impact on rights and duties, it would be necessary to assess the application of the specific regulation processing mechanisms.

Questions to address during the session

- What assurance mechanisms should be included?
- Are there any cases in which regulatory approval would be required?

References

2. Automated Administrative Actions – Law 40/2015 of 1 October on the legal system of the Public Sector – [link](#)

Art. 41.1 of the LRJSP (the Public Sector Legal System Act) defines an automated administrative action (AAA) as “**any action or task carried out entirely through electronic means by a public administration within the framework of an administrative procedure and in which a public employee has not directly intervened**”.

Each AAA must be approved within the scope of the relevant department or body through a resolution, which must establish the scope of the AAA and the body or bodies competent for the definition of the specifications, programming, maintenance, supervision and quality control and, if applicable, auditing of the information system and its source code. Likewise, the body that must be considered responsible for the purpose of appeal will be indicated.



A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

4. Floors – steps for the approval of algorithms

F5. Approval by resolution

References

2. The Catalan Tax Agency approves its AAAs by resolution – [link](#)

In resolution VEH/9/2020 of 9 January, the Catalan Tax Agency (ATC) approves a series of five automated actions which include the generation of certificates, resolutions, authentic copies and acknowledgements of submittal and payment, which do not require any manual intervention, and indicates the **conditions in which they can be carried out** and the bodies with which an appeal may be lodged if appropriate.

The resolution establishes the **signature system**: either being stamped by the ATC body or by secure verification code.

It identifies the **units responsible for the operation of the automated system**: for the procurement of the technology, the definition of the specifications and the auditing of the information system, and for supervising signatures.



3. Are algorithms regulations? – [Link](#)

The current legal system envisages that algorithms must be formally approved by the relevant entity through a resolution, although only when they involve no “human in the loop”. Andrés Boix, Lecturer in Law at the University of Valencia, argues that this mechanism is insufficient, since **the design and operation of algorithms affect the allocation of rights and duties**, like a regulation, and that they should therefore go through a **the same specific approval procedure that is used for regulations** and which guarantees all the mechanisms of **participation, publicity, control and procedural appeal** that are bestowed on regulations, assuming the necessary investment of time.

The US research centre [AI Now](#) recommends that algorithm development processes include public reporting and social participation mechanisms.





5. Attic floors: guarantees for algorithms already in use

A1. Transparency: algorithm data sheet



A2. Human oversight

A3. Questioning decisions


A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

5. Attic floors: guarantees for algorithms already in use

A1. Transparency: algorithm data sheet

 What does this step consist of?	 Questions to address during the session
<p>Transparency, through the publication of an algorithm data sheet which explains the problem being resolved, the risks identified, the decisions adopted and the data used, as well as the results of the bias analysis and global explicability techniques that have been applied.</p>	<ul style="list-style-type: none"> • What needs to be explained in the algorithm data sheets? • Does the source code need to be included in the data sheets so that it can be inspected? • Do the modifications made to the initial design need to be recorded?

References

<p>1. Amsterdam and Helsinki Algorithm Registers – link</p> <p>The city councils of Amsterdam and Helsinki have developed two twin portals to publish information on the algorithms they use. Let's take a look at the content for one of the algorithms, used for the purpose of reporting issues in the public space. This particular system consists of a web application through which residents can report incidents (uncollected rubbish, damaged street furniture, etc.), indicating the location and description and attaching photos. An algorithm processes the text of the description, classifies the incident and sends it to the relevant department, based on previous training with correctly classified previous incidents. The system makes it possible to greatly cut down the time needed to deal with such incidents.</p>	
---	--

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

5. Attic floors: guarantees for algorithms already in use

A1. Transparency: algorithm data sheet

References

The information in the register includes these sections:

Data sets, where the sources used to develop and operate the system, its content and its methods of use are indicated. The system described in the example was developed based on 300,000 old incidents, which cannot be published because they are free-text fields that could contain personal data. And the current system contains contact information on the people who report the incidents, which is not used in the algorithm and is erased once the incident has been resolved.

Data processing model, with a brief explanation of the logic used, together with a diagram and the source code. The performance score for the language processing algorithm used is also given, with a link to the study in which it was calculated.

Non-discrimination, where potentially discriminated groups and mitigation measures are indicated. In the example, the algorithm only recognizes Dutch and discrimination against other languages is not considered to be unfair. If unusual words (such as those from unusual dialects or language registers) are detected, the algorithm's service centre is informed and retrains the algorithm if deemed necessary to recognize them.

Human oversight, where both the mechanisms in which a human makes the final decision (human-in-the-loop) as well as those where automatic decisions are monitored (human-over-the-loop) are indicated. In the example, incidents where the classification is less than 40% certain are sent to the service centre for manual classification, while if an incident is sent to the wrong department, the department manually reclassifies it.

Informació més detallada sobre el sistema

Aquí podeu conèixer la informació que fa servir el sistema, la lògica de funcionament i el seu govern a les àrees que us interessin.

- Conjunts de dades Mostra més
- Processament de dades Mostra més
- No discriminació Mostra més
- Supervisió humana Mostra més
- Riscos Mostra més

Processament de dades Mostra menys

La lògica operativa del processament i raonament automàtic de dades realitzat pel sistema i els models utilitzat:

Arquitectura de models

YOLO (You Only Look Once), és una xarxa per a la detecció d'objectes. La tasca de detecció d'objectes consisteix a determinar la ubicació de la imatge on hi ha determinats objectes, així com classificar-los.

Paper original:
<https://arxiv.org/abs/1506.02640>

[Enllaç al codi font](#)

Contingut

Adjunt



Arquitectura de models

Imatge d'arquitectura de monitor d'un metre i mig

A3 - Questioning decisions
A2 - Human oversight
A1 - Transparency: algorithm data sheet
F5 - Approval by resolution
F4 - Analysis of bias and global explicability
F3 - Specifications
F2 - Data protection assessment
F1 - Socio-technical analysis
B3 - Proportionality
B2 - Caution
B1 - Awareness and training

5. Attic floors: guarantees for algorithms already in use

A3. Questioning decisions

 What does this step consist of?	 Questions to address during the session																						
<p>Possibility of questioning any decision made by the algorithm, with the help of channels to track them and local explicability techniques to interpret them.</p>	<ul style="list-style-type: none"> • Through which channel should the decisions be questioned? • Before whom should they be questioned? • How will decisions be explained? 																						
References																							
<p>1. Automated Administrative Actions – Law 40/2015 of 1 October on the legal system of the Public Sector – link</p> <p>The law establishes that Automated Administrative Actions must be approved by means of a resolution in which the body responsible for appeal purposes will be indicated.</p>																							
<p>2. French Commission on IT and liberties (CNIL) – link</p> <p>This data regulation body helps citizens to know and exercise their rights in the digital environment, such as not being the subject of automated decision-making or profiling; even if they have previously consented, the affected person can request human intervention from the relevant body, and if their request is not dealt with, they can file a complaint with the CNIL, whose stance against facial recognition has received worldwide attention.</p>																							
<p>3. Local explicability techniques such as LIME – link</p> <p>These approximation techniques make it possible to reconstruct which factors have had more weight in a specific decision made by an AI system, and can be used to justify the decision's degree of reasonableness in the event that it is questioned.</p>	 <p>NO INCIDENCIA / INCIDENCIA</p> <table border="1"> <caption>LIME Feature Weights</caption> <thead> <tr> <th>Feature</th> <th>Weight</th> </tr> </thead> <tbody> <tr><td>color</td><td>-0.18</td></tr> <tr><td>información</td><td>-0.15</td></tr> <tr><td>resumen</td><td>-0.12</td></tr> <tr><td>edad</td><td>-0.10</td></tr> <tr><td>relación</td><td>-0.08</td></tr> <tr><td>gpt</td><td>0.05</td></tr> <tr><td>información</td><td>0.03</td></tr> <tr><td>organización</td><td>0.02</td></tr> <tr><td>gpt</td><td>0.01</td></tr> <tr><td>edad</td><td>0.01</td></tr> </tbody> </table>	Feature	Weight	color	-0.18	información	-0.15	resumen	-0.12	edad	-0.10	relación	-0.08	gpt	0.05	información	0.03	organización	0.02	gpt	0.01	edad	0.01
Feature	Weight																						
color	-0.18																						
información	-0.15																						
resumen	-0.12																						
edad	-0.10																						
relación	-0.08																						
gpt	0.05																						
información	0.03																						
organización	0.02																						
gpt	0.01																						
edad	0.01																						