Generalitat de Catalunya
Departament de Polítiques Digitals
i Administració Pública
Direcció General
d'Administració Digital

Version 1

Data Ethics

Proposal for an Algorithm Ethics Framework

Office of Innovation and Digital Administration

2021

Contents

1. Creation of this document

The digital era, characterised by the availability of abundant data and the technology to process it, has enabled the development of advanced services that can help the Public Administration in the exercise of its duties and in overcoming the challenges it faces in order to effectively serve the public. In this regard, the Administration of the Generalitat de Catalunya has worked especially hard during the 12th term of office on introducing proactive and personalised public services. However, the implementation of new and powerful service provision mechanisms opens up the possibility of knock-on effects that are not sufficiently well known and that could potentially have an impact on aspects such as fairness or privacy. In light of this, it is advisable to begin considering the precautions and limits to be taken into account in the implementation of these services; limits not only in the form of regulations that need to be kept constantly up-to-date, but that must incorporate the perspective of ethics.

The Administration of the Generalitat de Catalunya has launched a participatory process to begin considering these aspects in the form of the ApLab learning laboratories. With representatives from a range of specialist fields (from law to technology and a variety of other functional areas) from all ministries and with the inclusion of external experts and representatives from civil society, the following sessions have been held:

- ApLab on “Ethics of proactive and personalised services”, 30 September 2020.
- ApLab on “Algorithm approval”, 10 November 2020.
- ApLab on “Ethics of algorithms”, 19 January 2021.

This document is the final product of the work carried out through the dossiers that have been created for each session and the contributions and deliberations that have taken place.

2. The ethical impact of proactive and personalised public services

The ethics of algorithms is not an established discipline in public administrations, although some have begun to make headway in this area. In the case of the Administration of the Generalitat de Catalunya, the process has been initiated as a result of the interest in implementing proactive and personalised services in a way that is ethically responsible. As these services are being developed thanks to the possibility of implementing them through algorithm-based systems, ethical reflection on proactive and personalised services directly implies the need to address the ethics of algorithmic systems.

To give a brief definition of the concept, we can say that proactive services are provided when the recipient is known to need them, while personalised services are provided in the way they are needed, and together, proactive and personalised public services are those provided by the Administration based on its knowledge of each user. Proactivity is multilayered and does not necessarily require complex algorithms, although in the digital age, the availability of big data and automated decision-making systems such as artificial intelligence (AI) is allowing it to become more extensive.

The analysis of proactive and personalised services from an ethical perspective identifies a number of potentialities and risks:

- They can help the Administration to address challenges it must respond to, but at the same time they can produce knock-on effects. For this reason, it is important to approach them from the perspective of ethics, not understood as a simplistic distinction between right and wrong practices, but as a reflection on complexity and on the risks that can be taken.
- They can contribute to social justice to the extent that they help people to exercise their rights. But public services may contain biases for or against certain groups, such as the digital divide, and the implementation of these services through AI algorithms makes it difficult to detect these biases. To ensure that digital services are fair, they need to be designed with an omni-channel perspective, and algorithms need to be transparent and explainable so that they can be analysed, especially in sensitive areas. Clear boundaries and strict and transparent controls are needed.
- They can help people to gain autonomy and to focus their energies where they really want. Personal autonomy and dignity are not threatened by proactive and personalised public services, as long as it is guaranteed that people can decide at all times whether or not to receive them, with consent that need not always be written, but must be the result of a well-informed choice. In both the public and private sectors, personal data governance is needed to ensure individual control of personal data, although by following data protection rules privacy can be protected without giving up services that offer undoubted benefits. In this respect, the data governance model of the Administration of the Generalitat1 will guarantee, among other things, that individuals have access to information on what data the Administration has about them, how it uses them and the corresponding processing, and will implement the necessary policies in areas such as data quality and security.

Ultimately, there are risks in the implementation of proactive and personalised public services, for example the fact that AI is still immature for certain uses, for which reason the Administration should regulate the validation of algorithms. However, knock-on effects can be avoided, and where new forms of service delivery outperform traditional ones, they should be encouraged.

[1] Articles 10 and 11 of Decree 76/2020 of 4 August on digital administration.

3. Structure of the framework

The implementation of advanced algorithm-based public services requires a number of safeguards or precautions that may constitute an approval procedure, i.e. a sequence of steps that must be taken before an algorithmic system can be considered as offering sufficient guarantees to be implemented.

However, there is still a need for guarantees after the algorithm has been approved, as both the algorithm and its environment may change over time and therefore need to be monitored, and citizens – who may exercise their civic right to know and question the functioning of the algorithms applied to them – need to be included in this monitoring.

Finally, both the *ex-ante* approval procedure and the *ex-post* guarantees do not exist in a vacuum, but require a substratum of organisational principles and elements to make them possible.

For all these reasons, it is proposed to establish an algorithms ethics framework in the Administration of the Generalitat de Catalunya, consisting of the following elements:

1. Principles
    - Awareness and training
    - Caution
    - Proportionality
2. Guarantees prior to the implementation of algorithmic systems
    - Socio-technical analysis
    - Data protection assessment
    - Procurement specifications
    - Analysis of bias and global explainability
    - Approval by means of a decision
3. Guarantees on algorithmic systems already in operation
    - Transparency
    - Human oversight
    - Questioning decisions

4. Principles

4.1 Awareness and training

Mitigating the risks of algorithmic systems requires, first and foremost, a basic knowledge of these systems and an awareness of their potential impacts. It is therefore necessary to raise awareness among the wide range of stakeholders involved in the decision-making, design, implementation, operation and monitoring of algorithm-based systems.

Awareness of algorithmic systems and their risks must be spread through the usual means of dissemination and awareness-raising, such as the media and social media, by means of concrete examples:

- Among the public, since their understanding of the issue is essential if an effective algorithm transparency policy is to be achieved.
- Especially among young people and children, who have already been born into the digital environment and may therefore require a broader perspective to help them judge it critically, e.g. with regard to aspects such as the transfer of their data.
- In particular, among the people who use each of the public services that involve advanced data use, so that they can adopt knowledge-based attitudes and decisions in their relationship with the services.

Through the Public Administration School of Catalonia, it is also necessary to raise awareness among the following public servants:

- Senior management, in their role as high-level decision-makers, through the induction handbook and refresher training.
- Middle management and professionals in fields such as law and IT, who are involved in the detailed decisions related to the algorithmic system design and implementation process.
- Managers of procedures involving algorithms, who should be aware of the potential ethical impacts of the algorithms they work with.

Finally, public employees must be trained with the skills needed to carry out specific tasks that are part of the present framework, such as:

- Drawing up technical specifications for algorithmic systems.
- The preparation of analyses.
- Validation of analyses carried out by third parties.
- Interpretation of analysis results and the adoption of mitigating measures.

4.2 Caution

The Public Administration has a lot of room for digital transformation through established technologies, for example to help it overcome the challenge of interoperability by organising data to make decisions based on simple rules. For critical uses, security must be prioritised over immature technologies that may produce risks that are difficult to assume. On the other hand, less critical uses can be employed as an opportunity to gain experience with more innovative technologies which could be in widespread use in the future.

It is difficult to distinguish the more prudent from the less prudent uses by simple rules that limit certain technologies or their use to certain sectoral areas, since each of these areas and technologies encompass both high-risk and low-risk cases.

It is therefore necessary to adopt the precautionary principle, which states that we should not introduce all technically possible innovations, but only those which have an acceptable and mitigable impact, applied through a risk-based approach. This principle, which has underpinned personal data protection since the adoption of the General Data Protection Regulation (GDPR), should therefore be extended to the other risks associated with the implementation of algorithmic systems.

4.3 Proportionality

The controls to be applied in the implementation of algorithms must be graduated in proportion to the risk involved in each case, understood as the product of the probability of the algorithmic system malfunctioning and the possible impact. This assessment will determine where controls need to be implemented (*ex-ante* audit, *ex-post* inspections, etc.) and the obligations of both providers and the Administration of the Generalitat.

The following factors should be taken into account in assessing the likelihood of malfunction:

- Complexity of the algorithm. Classical algorithms that respond to an exact rule are much more predictable than machine learning algorithms, and among the latter, deep learning algorithms are particularly inscrutable due to their complex structure.
- Previous certifications or validations of the algorithm in other areas of the public sector.
- Depth and quality of data. An algorithm trained with insufficient data or with high margins of error will be inaccurate.

The impact of malfunctioning should be assessed in terms of elements such as the following:

- Type of administrative action: regulation, promotion, policing, etc.
- Economic cost linked to the administrative action.
- Situation of the target group.

- Use of specially protected data.
- Definition of profiles.

For the purpose of ensuring objectivity, the risk level of an algorithmic system should be determined by an interdisciplinary and interdepartmental team with members from the following areas:

- The functional unit responsible for the algorithmic system.
- Public employees with expertise in ethics, cybersecurity, law, ICT and data protection.
- External experts and representatives of civil society.

5. Guarantees prior to the implementation of algorithmic systems

5.1 Socio-technical analysis

Before developing or acquiring an algorithm, it is necessary to analyse the approach taken to transferring a complex social problem to data processing. This analysis should identify the main risks of this transfer and the social groups that could be negatively affected and should be protected. As a result of this analysis, the required mitigation strategies should be proposed or, if necessary, the algorithmic system should be redesigned or not implemented.

This analysis should be carried out by an interdisciplinary team to ensure an understanding of all the relevant aspects:

- Of the problem the system is intended to solve, and the sector in which it operates:
  - managers and employees of the unit responsible
- Of the means used to resolve it:
  - data science professionals
  - ICT professionals.
- Of the people potentially affected by the service:
  - sociology professionals
  - public employees directly serving the people to whom the system is addressed
  - representatives from among the targeted users.
- Of cross-cutting criteria that must be met by algorithmic systems:
  - legal
  - data protection
  - cybersecurity.

The Administration of the Generalitat will obtain the professional profiles needed to form these interdisciplinary teams, by means of

- training public employees
- hiring new professionals
- collaboration with universities
- and the voluntary participation of citizens.

5.2 Data protection assessment

European data protection legislation establishes two instruments for analysing the risk of a system to personal data and for the design of mitigating measures: the risk level assessment and, in the event of high risk, the data protection impact assessment. The corresponding instrument must be used at the beginning of the algorithm design process.

According to the principle of data minimisation, it is necessary to identify what data are needed and whether the risk of including them is justifiable and safeguards can be put in place to protect them. However, in doing so, it is necessary to consider not only what data are needed for the algorithm to work, but also what data need to be collected to ascertain whether the algorithm is acting fairly or is biased, in order to monitor and mitigate indirect discrimination, which is where the algorithm ends up producing results that discriminate against a certain group based on variables that indirectly target that group. The processing must therefore include the data that allow identification of the groups to be protected, even if they are not used by the algorithm.

Another issue to be debated on the use of algorithms in the activities of the Administration is the fit between automated individual decision-making, regulated under Article 22 of the GDPR, and the automated administrative actions provided for under Article 41 of Law 40/2015 of 1 October on the legal regime of the public sector. The GDPR establishes the right of individuals not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or significantly affects them in a similar way. In the case of such processing, legal authorisation or consent to the processing of individuals is required.

In relation to the requirements for considering an automated administrative action as being automated individual decision-making, there is a rich legal debate on what constitutes an act of decision-making. In this respect, it is necessary to distinguish between classical algorithms and algorithms based on artificial intelligence. In the former, which are limited to the exact application of the agreed and published rules, the application is merely an auxiliary element that the human operator must take into account (or not) when determining the content of the administrative action. As for the second category, based on artificial intelligence, these are algorithms that, through training, learn to make judgements on which to base a decision that is not provided for by the legal system. In this case, and if the algorithm does not simply issue a pre-programmed decision but generates an action with an unforeseen outcome, it can be clearly stated that it is the algorithm that makes the decision, so that the safeguards provided for in the GDPR with regard to the processing of this information will have to be applied. It must therefore be concluded that not all administrative actions issued on the basis of the results provided by the algorithm are automated decisions in accordance with the requirements of the GDPR.

In any case, providing individuals with control over their data is essential. In the Administration of the Generalitat, the Consent Register[2] will allow citizens to control the proactive and personalised services they wish to receive and to know what data is used and what processing is applied, including automated decisions.

Together with the legal perspective on data protection there must also be an ethical perspective, which in the Administration of the Generalitat will be led by the Data Ethics Committee, a space for reflection that brings together the internal perspective of the Administration and the participation of civil society groups and external experts.

5.3 Analysis of bias and global explainability

Before developing an artificial intelligence algorithm, it is necessary to analyse whether the data available to train it adequately represent the various groups to be protected. The variables on which to measure potential biases will be those identified in the socio-technical analysis.

Once the algorithm has been developed, and before putting it into production, the fairness of the algorithm's response to the dataset used for validation must also be analysed, since the algorithm's response will depend not only on training, but also on design, and therefore it may be the case that an algorithm makes unfair discriminations despite having been trained with unbiased data.

If computationally feasible, it will also be necessary to characterise the criteria by which the algorithm makes decisions using global explainability techniques to assess whether these criteria fall within what is functionally considered fair and reasonable.

The algorithm's development and maintenance provider shall perform the analyses and deliver them in the format specified by the Administration of the Generalitat. These analyses will be received by professionals from the Administration of the Generalitat with sufficient knowledge of statistics and with the support of automatic tools that make it possible to verify a minimum level of quality in the analyses delivered. The ministries will have functional managers who are trained to interpret the results of the analysis and request mitigating measures.

[2] Article 31.2 of the Digital Administration Decree.

5.4 Technical specifications

The provisioning of algorithms must adhere to algorithmic fairness and transparency criteria. These criteria must be included in the technical specifications, or in the requirements documents in the case of procurement mechanisms that do not involve the drafting of specifications, such as the commissioning of evolutionary improvements within the framework of a pre-existing application maintenance contract.

Companies contracted to develop and maintain artificial intelligence algorithms should be able to provide, *inter alia*:

- the analysis of bias in the data to be used to train the algorithm
- the analysis of bias in the results of the algorithm using the validation dataset, once the algorithm has been developed and trained and prior to its implementation
- periodic analyses of bias in the real-world performance of the algorithm
- the analysis of the criteria applied by the algorithm using global and local explainability techniques.

A multi-departmental interdisciplinary team will provide the ministries with a guide for drawing up technical specifications. The members of this team should be available to the contract awarding committees to advise them on the assessment of compliance with the conditions of the bids submitted in tenders involving algorithms.

5.5 Approval of automated administrative actions

Article 41.1 of Law 40/2015 defines automated administrative actions and establishes two essential elements:

- The action must take place within the framework of an administrative procedure.
- The preparation of the action must not directly involve a public employee, i.e. it must be carried out entirely by automatic means.

Regulation is needed prior to the establishment of these actions, including:

- The definition of specifications, programming, maintenance, supervision and quality control.
- Auditing of the information system and its source code.
- The body responsible for the purpose of appeals.

Within the sphere of the Administration of the Generalitat, this prior regulation must be carried out by means of a **decision of the competent body, which must be issued prior**

to the launch of the service[3]**.** The decision should incorporate the definition of the specifications, programming, maintenance, monitoring and quality control and, where appropriate, auditing of the information system and its source code[4]. It should also indicate the body responsible in the event that the automated action is challenged. Furthermore, the resolutions must be published on the Electronic Office of the Administration of the Generalitat[5] in order to guarantee public awareness of these aspects.

6. Guarantees for algorithms already in use

6.1 Transparency: algorithm data sheet

The Administration of the Generalitat de Catalunya, following the path taken by the city councils of Amsterdam and Helsinki, should have a portal with easily understandable data sheets on the algorithms it uses which explain 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 explainability techniques that have been applied.

The data sheets should include easily understandable information on the following points:

- The problem the algorithmic system has been designed to solve.
- The logic used to approach the task and the justification for the decisions that have been taken.
- The data used to train and validate the algorithm: where they have been taken from and under what rights.
- The criteria applied by the system.
- The link to the source code, when public.
- The risks that have been identified and biases that have been measured.
- The measures that have been implemented against risks and biases and, in particular, the type of human oversight that has been introduced.
- The body to which appeals against the algorithm should be addressed.

For this policy of transparency to be effective:

- The data sheets should be drafted on the basis of a common framework or glossary so it is easier for citizens to understand and assess them.
- In this regard, it would be very useful to introduce a system of quality seals or labels based on expert assessment to characterise algorithmic risk, similar to what has been done in other areas such as energy efficiency in buildings.

[3] Article 54 of the Digital Administration Decree.

[4] Article 41.2 of Law 40/2015

[5] Article 54.2 of the Digital Administration Decree.

- The data sheets should be kept up to date during the life cycle of the algorithm and should include any variations in the algorithm, the results of periodic inspections and indicators characterising the results of the algorithm.
- As well as this general transparency of the Administration towards citizens, the Administration must also offer personalised transparency to each citizen, who must know which algorithms are using their data and must be able to control this use in the cases envisaged by law. In the Administration of the Generalitat de Catalunya, it is necessary to exploit the possibilities offered in this regard by the Consent Register, in accordance with the provisions of the Decree on Digital Administration.[6]

6.2 Human oversight

The risk introduced by algorithmic systems, and especially those with a factor of uncertainty, such as AI-based algorithms, can be mitigated by pairing them with human supervision. This measure should be guided by the principle of proportionality and should therefore be applied more intensively in higher-risk cases.

Thus, in cases where the system has a higher impact, or where its accuracy is lower, the algorithm may simply recommend an action to a human decision-maker who will confirm or alter the recommended action at his or her own discretion (human-in-the-loop). In cases of lower risk, the algorithm can be left to make decisions directly, with *ex-post* supervision of correct functioning (human-over-the-loop).

Supervision may vary during the life cycle of the algorithmic system. Direct monitoring is particularly important in the early stages of implementation of a technology, although later on a switch to periodic or sample monitoring can be considered.

Just as the limitations of algorithms must be taken into account, the limitations of humans must also be borne in mind. Repetitiveness of tasks and lack of time or judgement can degrade the quality of human supervision, so appropriate and proportionate design is necessary. Moreover, the biases of public employees could combine with those of the algorithm, so training should be given that is specifically designed to avoid such biases.

Ex-post (human-over-the-loop) supervision should be carried out by staff with a good knowledge of the specific sector but who are organisationally distanced from the unit responsible for the algorithmic system, so that they can question it from a position of neutrality. In addition, the algorithm's functioning and results should be transparent so that expert individuals and organisations from civil society can take part in this supervision.

[6] Final provision five of Decree 76/2020 of 4 August on digital administration.

6.3 Questioning decisions

Citizens have the right to legally challenge the decisions of algorithmic systems in two ways:

- by means of an appeal against an individual administrative action produced by the system
- by challenging the algorithmic system before the competent body, which in the case of automated administrative actions shall be specified in the decision approving the system.

It is also necessary to consider ways that algorithmic systems not covered by the regulation governing the automated administrative action can be legally challenged, such as through the periodic passing of a law on algorithms.

In anticipation of possible appeals or challenges, bodies that set up AI-based algorithmic systems will have to be able to explain:

- the reasons for any particular decision, for which they can use local explainability techniques
- the general criteria applied by the system, for which they may use global explainability techniques.

7. Final considerations

The implementation of advanced algorithm-based public services both in the private and public spheres has generated concerns that have begun to gain weight in a matter of little more than months. Very few public institutions have established spaces for deliberation and precautionary measures to mitigate the risks arising from algorithms. With this proposed framework, the Administration of the Generalitat de Catalunya joins the ranks of those administrations committed to the responsible use of algorithms, fully aware that the measures proposed will not constitute a definitive guarantee, but only a first line of defence that will need to be perfected with practice and with an increase in possible uses of data.

Parallel to the process that has led to the drafting of this proposed framework, the creation of the Data Ethics Committee has been approved as a consultative interdepartmental body at the service of the Administration of the Generalitat de Catalunya and its institutional public sector. The purpose of the Committee is to guarantee a space for ethical reflection on the deployment of the Administration's data model that helps the subjects of the Administration of the Generalitat to make decisions, generate knowledge and guidance, develop advanced data uses and generate good practices and attitudes to guide the development of the digital Administration. This Committee is the ideal instrument for taking into consideration the proposals of this framework and for promoting its implementation, revision and continuous improvement.