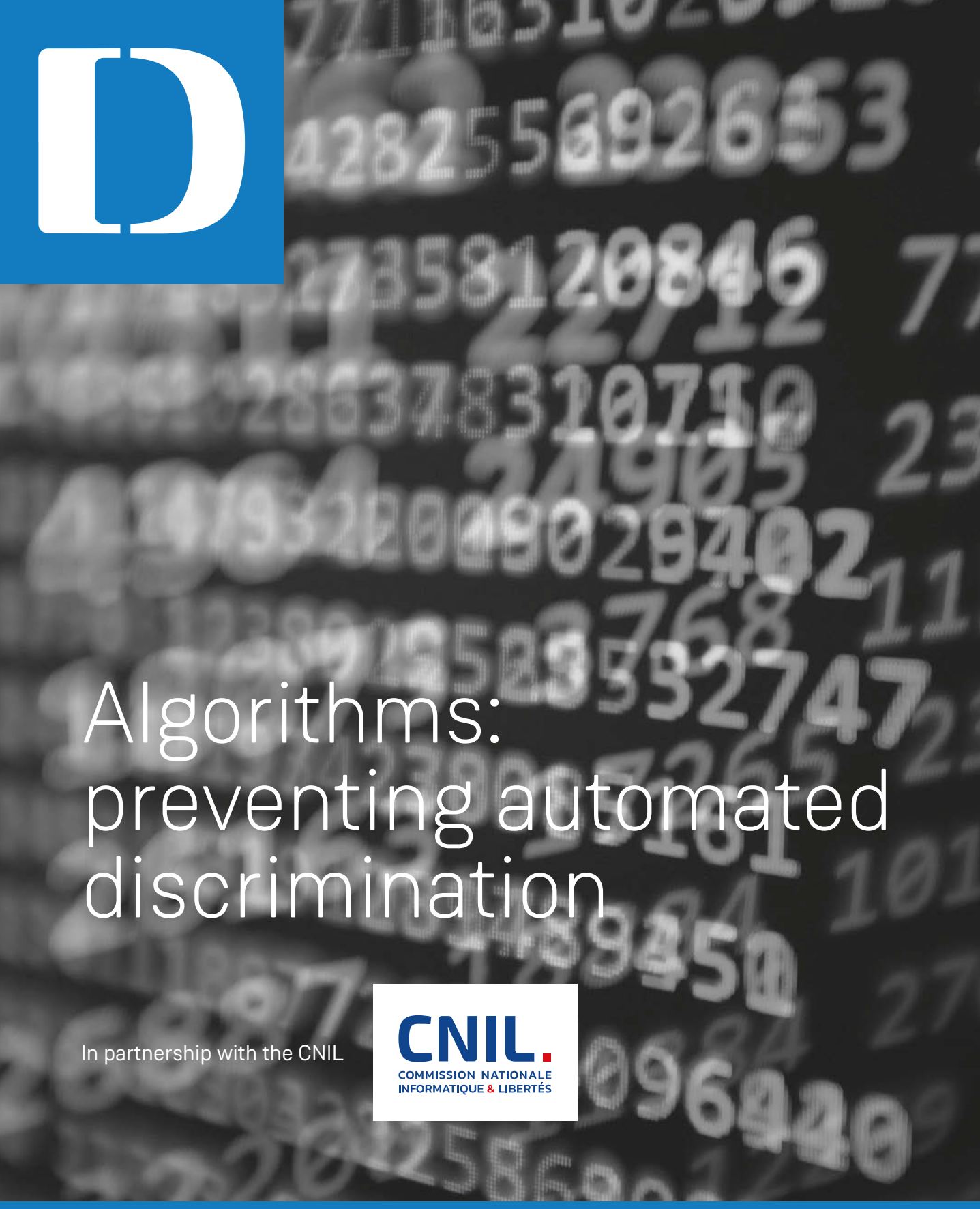

# Algorithms: preventing automated discrimination

In partnership with the CNIL

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

**When it comes to the law, we are all equal**

Défenseurdesdroits
RÉPUBLIQUE FRANÇAISE

# Algorithms: preventing automated discrimination

# Introductory remarks

**During the current global health crisis, the use of digital tools has increased and diversified as never before, resulting in major debates. These digital tools are often based on algorithms, although users are not always aware or informed.**

Recourse to algorithms as a basis for public or private decision-making is not a new phenomenon: the automated calculation of financial risk performed by banks ("*scoring*") and which involves combining various criteria drawn from information provided by loan applicants has become more widespread over the last decades. Yet, as noted by the Conseil d'Etat, intensive use of algorithms as a result of computers' new calculation power and the mass processing of what is now a large amount of data marks an "unprecedented turning point"[1].

In just a few years, the use of algorithms has expanded to the private sector and to administrations[2]. Today, such processes can be found in fields that are as essential to individuals as access to social benefits[3], policing and justice[4], the running of organisations such as hospitals, access to public services and recruitment procedures[5].

Since 2006, machine learning technologies have taken off. Once rolled out, these learning systems continue to evolve, striving for perfection.

These technological evolutions, which are still in progress, are undeniable sources of progress for individuals and society, allowing for quicker, more reliable and personalised results as well as new analyses in many fields.

However, **the Data Protection Commission-CNIL and the Defender of Rights have both, in their own area of expertise, voiced their concerns regarding the impact of these algorithmic systems on fundamental rights**[6].

It is with this mindset that the Defender of Rights is acting, in partnership with the CNIL, in the hope of highlighting **the considerable risk of discrimination that each and every one of us is exposed to by the exponential use of algorithms in all aspects of our life**.

**This topic has long been a blind spot in public debate. This must change**.

[1] Conseil d'Etat, *Puissance publique et plateformes numériques : accompagner « l'ubérisation »*, La documentation française, 2017, p. 59.

[2] See, for example: The State's Inter-ministerial Directorate for Digital, Information and Communications Systems (Direction interministérielle du numérique et du système d'information et de communication de l'Etat), DINSIC, Guide des algorithmes publics 2019.

[3] National Delegation to Combat Fraud (Délégation Nationale à la Lutte contre la Fraude), *Le « data mining », une démarche pour améliorer le ciblage des contrôles*, Paris, 14 January 2014.

[4] Soraya Amrani Mekki, "Justice prédictive et accès au juge", La Justice Prédictive, Actes du Colloque of 12 February 2018 organised by Conseil d'Etat and Cour de cassation Lawyers Council for its bicentenary in partnership with the Paris-Dauphine PSL University, Paris, Dalloz, 2018.

[5] Christine Bargain, Marie Beaurepaire, Dorothée Prud'homme, *Recruter avec des algorithmes ? Usages, opportunités et risques*, AFMD, 2019.

[6] CNIL, Travaux sur le système APB (decision no. 2017-053 of 30 August 2017); *Comment permettre à l'Homme de garder la main ?* Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle, 15 December 2017. Defender of Rights, Guide - Recruter avec des outils numériques sans discriminer published in 2015, Opinion no. 15-25 of 1 December 2015 on security in stations; *Report titled "Lutte contre la fraude aux prestations sociales : à quel prix pour les droits des usagers ?"*, September 2017, Parcoursup decisions (2018-323 of 21 December 2018 and 2019-21 of 18 January 2019), Opinion 18-26 of 31 October 2018 on the Draft Programming and Reform Act for Justice, opinion 19-11 of 5 September 2019 on the Draft Act on Bioethics.

# How can algorithms be discriminatory?

At first glance, algorithms sort, categorise and organise information by eliminating any prejudice and bias specific to human beings. Thus, they should be able to ensure the equal treatment expected by applying the same criteria and weighting regardless of the requester's original or sexual orientation for example.

In reality though, there is no technological magic or mathematical neutrality: algorithms are designed by humans using data that mirror human practices. As such, **bias can be introduced into every stage of the development and deployment of systems: as from the intention that initially governs the algorithm's development, during the creation of the computer code, the executable code, during execution, in the context of execution and maintenance**[7].

Some completely intentional bias can also result from the inclusion of prohibited grounds for discrimination in an algorithm. Some reasons can be taken into account to justify the criteria used by an algorithm in some specific cases such as state of health for insurance, age for bank loans or the place of residence to adjust premiums, if their use is considered proportionate to a legitimate purpose[8]. However, criteria such as gender or origin cannot constitute lawful criteria, regardless of context.

Nevertheless, the discriminatory effects of algorithms are often based on mechanisms that are less visible than the inclusion of easily-identifiable prohibited grounds for discrimination in the algorithm.

## Biased data

**Discriminatory mechanisms are frequently based on the bias of the data selected and used by a traditional algorithm or fed into a learning algorithm during its learning phase and after**.

One of the most frequent biases is based on a **lack of representativity in the data used**. For example, in 2018, a study explained why some facial recognition systems, which are based on learning techniques[9], found it harder to identify women, people who are not white and more so women of colour, by generating a high error rate for these populations: the datasets that this model was based on was characterised by a large predominance of male white faces[10]. The issue is similar for voice identification technologies: having not been designed with women and their voices in mind, and not having been built (and therefore fed with "female" data) and tested in this regard, the system does not work as well for women[11].

The data integrated into algorithmic systems or used to teach a machine learning system can also be biased when they are the **mathematical result of past often-discriminatory practices and behaviour and of systemic discrimination present in society**.

[7] Barocas S., Selbst and Andrew D. "Big data's disparate impact", California Law Review, June 2016 Vol. 104, no. 3, pp.671-732.

[8] C.E., 30 October 2001, no. 204909, *association française des Stés financières*. See the article "Testing, scoring, ranking...", Revue trimestrielle de droit civil, July-September 2002, no. 3, p. 498.

[9] CNIL, Reconnaissance faciale. Pour un débat à la hauteur des enjeux, 15 November 2019.

[10] According to MIT researcher Joy Buolamwini's study, the error rates of Amazon's Rekognition software were 1% for lighter-skinned men, 7% for lighter-skinned women, 12% for darker skinned men and 35% for darker skinned women. See Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." *MIT News*, 11 February 2018.

[11] During an internship focused on voice recognition for helicopter pilots, when a woman was put at the commands, the system did not work as well, representing a serious safety issue. (TUAL M., "La diversité humaine est un enjeu central pour le développement de l'intelligence artificielle", *Le Monde*, 30/07/2018).

In available employment data, women are less well represented and tend to occupy certain business sectors and lower positions with lower pay. Based on such data, an algorithm might deduce that women are not as productive as men and do not reach positions of responsibility. As a result, an algorithm used for recruitment based on biased data will reproduce such biases, and even exacerbate them[12].

# False neutrality of algorithms and true discriminatory effects

**The use of apparently neutral criteria, i.e. criteria that does not include prohibited grounds for discrimination, can have discriminatory effects** as highlighted by the Defender of Rights in its Parcoursup decision[13]. In this case, university algorithms which take into account the seemingly neutral criteria of institution of origin could, indirectly, result in discriminating against youths of foreign origin, given the strong residential and educational segregation observed in Ile-de-France in particular.

**Most often, these discriminatory effects are caused by the combination of several neutral criteria**. The criteria and data in question may even seem far removed from prohibited reasons, but their correlation provides similar results to those which would have been obtained had the protected characteristic been applied. Learning algorithms, and the many correlations they make between massive amounts of data, can easily generate such effects. In this case, belonging to a protected category is encoded in "neutral" data.

Designed to maximise its ability to find similar characteristics among massive amounts of data, the programme recreates a whole matching the protected category, and applies specific processing to it.

In order to target its advertising, the American supermarket company Target developed a predictive model to identify pregnant clients based on their purchase habits concerning 25 products[14]. Such models could be used for discriminatory purposes or could have discriminatory effects.

**Algorithms may combine several sources of bias and ruin even the best intentions**. Several American studies have recently demonstrated the discriminatory nature of the main "smart" system models used to automatically detect hate speech for moderation purposes: the probability of having one's message reported as offensive or hate speech by the system was 1.5 times higher among Afro-American internet users. Such biases come from learning data: the panel of data was created by humans who first classed the messages containing abusive language as offensive or hate speech. These biases are also exacerbated by the technical limits of the system which struggles to identify the nuances of a language and put slang or sarcastic statements in their context for example[15].

[12] EEOC, Conference - Big Data in the Workplace: Examining Implications for Equal Employment Opportunity Law, 13 October 2016.

[13] Decision 2019-021 of 18 January 2019 on the operation of the national platform for pre-registration for the first year of higher education (Parcoursup).

[14] See KIM PT, "Data-driven discrimination at work", 58 Wm. and Mary Law Review 857, 2016.

[15] "The algorithms that detect hate speech online are biased against black people", 15 August 2019, vox.com.

# Invisible and potentially massive discrimination

**The discriminatory effects of algorithms can often only be measured by researchers at group level. They risk remaining completely invisible for victims**.

Furthermore, while the cognitive biases of one human being vary depending on circumstances and contingently translate into discriminatory practices, **the discriminatory biases integrated by an algorithm are applied automatically and could systematise discrimination**. There is a significant risk of reinforcing "essentialization" and "stereotypes" as the algorithm's predictive nature is based on the behaviour or homogenised characteristics of groups of people. These systems therefore could "reinforce discrimination and prejudices by giving them an appearance of objectivity"[16].

While the discriminatory effects of the algorithm are not always identifiable at individual level, the seemingly neutral algorithmic system may result in discrimination against protected social groups, which could translate, for example, into lesser access to the goods sought or a higher error rate produced by the system in their regard. This risk of discrimination is even greater for social groups having already been the victim of major systemic discrimination in society, for example women, people with a disability or immigrants.

By integrating former discriminatory practices as part of a dataset used for its learning phase, **the bias of "smart" systems tends to increase as they are rolled out**.

Predpol software enables many police forces to direct their action and "rationalise" their activity by identifying "hot points" where there is a higher risk of offences being committed, in order to increase patrols. This model also takes accounts of influence factors such as population density, the proximity of bars or means of transport. However, the predominance of information on the places where past offences and crimes have been committed is problematic. In the United States as in other countries, police controls, arrests and places where they decide to patrol target minorities and certain areas much more than others. Based on Predpol's suggestions, police forces would be mainly directed to these districts and would observe new offences, thereby feeding the learning base with new biased data. Algorithms could thus cause feedback loops in which stereotypes, discrimination and inequality mutually reinforce one another and contribute towards the long-term crystallisation of situations of inequality[17]. **Only by precisely and regularly checking the learning algorithm's results can it be ensured that the algorithm does not become discriminatory over the course of its successive encoding**.

Lastly, it should be added that **these systems tend to target and control, and therefore stigmatise, members of already-underprivileged and dominated social groups more than others**[18]. In 2019, several associations brought legal action against the Dutch State to have an algorithm developed by the Ministry of Social Affairs and Employment to predict the likelihood of an individual committing benefit and tax fraud declared unlawful.

[16] Dunja Mijatovic, Commissioner for Human Rights, "Safeguarding human rights in the era of artificial intelligence", Commissioner for Human Rights Comment, Strasbourg, 3 July 2018.

[17] Hiring by Algorithm: predicting and Preventing disparate impact - Ifeoma Ajunwa, Sorelle Freidler, Carlos Scheidegger, Suresh Venkatasubramanian; Draft of January 2016.

[18] Virginia Eubanks, *Automating inequalities. How High-tech tools profiles, police, and punish the Poor*; St. Martin's Press, January 2018.

During the hearing, the government acknowledged that this algorithm targeted districts containing a higher number of social benefit recipients, despite the lack of evidence that these districts showed a higher benefit fraud rates[19].

# Recommendations

The right to non-discrimination must be effectively respected under all circumstances, including when a decision involves recourse to an algorithm.

The extensive use of algorithms is - in the words of Cathy O'Neil - a "weapon of Math destruction" as regards equality issues[20]. Nevertheless, despite the first alarm bells rang by the Villani report[21] and a few initiatives[22], awareness is slow to emerge in France: **algorithm designers, like the organisations buying and using these types of systems, do not demonstrate the necessary vigilance to avoid a type of invisible automated discrimination**.

Yet, the fairness principle, which poses the notion of "users' interests" as an obligation for the person responsible for the algorithm, like the principle of vigilance and reflexivity which involves regular, methodical and deliberative checks on learning objects, should guide reflection and action[23].

It should be reminded that **non-discrimination is not an option but is part of a legal framework** which sets out an analysis grid to identify situations of unequal treatment in order to implement a fundamental right: that to not be discriminated.

Organisations using algorithms cannot escape their responsibilities under the cover of ignorance, technological incompetence or opaque systems. **Algorithmic biases must be able to be identified and corrected and those responsible for discriminatory decisions as a result of algorithmic processing must be sanctionable**.

As highlighted by existing literature, the lack of transparency of the systems implemented and the data correlations enabled by algorithms, often entirely invisibly, render the protection offered by law uncertain and even ineffective.

Thus, how can one exercise a right to recourse when one is not even aware of being the victim of discrimination as a result of an algorithm, when the organisation using the algorithm itself is not aware of it, when the designer of the algorithm will not or cannot explain how such a tool works? How can one find out whether an algorithm is discriminating a given social group? And, if such is the case, how can these breaches to rights be sanctioned? The work carried out alongside our European counterparts as members of the Equinet network[24], such as the cross-disciplinary seminar on "Algorithms, bias and combatting discrimination" organised on 28 and 29 May 2020 in partnership with the CNIL,

[19] Open Democracy, "Welfare surveillance on trial in the Netherlands", 8 November 2019. The Hague Court issued a decision on 5 February 2020 acknowledging that the government had breached the right to privacy and family life set out in Article 8 of the ECHR and ordered that it cease using this algorithm. The judges based their decision on the fact that the algorithm Syri lacked transparency. The court did not address a possible breach of Article 22 of the GDPR which bans automated decision-making in some cases.

[20] Cathy O'Neil, *Algorithmes. La bombe à retardement*, Les Arènes, 2018 (USA, 2016).

[21] Cédric Villani, *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*, Report to the Government, 8 March 2018.

[22] Telecom Paris Tech, *Algorithmes: biais, discrimination et équité*, February 2019; Aude Bernheim, Flora Vincent, *L'intelligence artificielle, pas sans elles !*, Laboratoire de l'égalité, Belin editions, 2019; Institut Montaigne, *Rapport Algorithmes : contrôle des biais SVP*, March 2020; Collective report ordered by the Etalab mission, *Ethique et responsabilité des algorithmes publics*, ENA, Class of 2018-2019 "Molière", June 2019.

[23] 40th International Conference of Data Protection & Privacy Commissioners, Declaration on Ethics and Data Protection, 23 October 2018.

[24] Equinet, *Regulating for an equal AI: A New Role for Equality Bodies. Meeting the new challenges to equality and non-discrimination from increased digitisation and the use of Artificial Intelligence*, June 2020.

has highlighted the lack of legal and technical expertise and the need to devise effective countermeasures. The guidelines published by the European Commission in April 2019 provide indications[25] and European experts' first conclusions[26] call for mobilisation to match the stakes.

Without in-depth reflection and mobilisation by public authorities, there is a significant risk in France that the right to non-discrimination will not be able to fulfil its purpose and protect the population.

As part of its mission to combat discrimination and promote equality, the Defender of Rights therefore wishes to raise awareness, in partnership with the CNIL, of the **need to mobilise as from today to prevent and correct such discrimination**.

While awaiting such a mobilisation, which the Defender of Rights intends to fully participate in over the coming months, the following aims should contribute towards launching and structuring a necessary collective reflection.

## Inform and raise awareness amongst professionals

The social reality of discrimination and the framework of anti-discrimination law are still little known and rarely considered by data and algorithm experts in Europe. There are significant acculturation and training issues, with IT and data analysis professions - which are often criticised for lacking diversity[27] - being still too unaware of the risks to fundamental rights caused by algorithms.

Reciprocally, professionals who purchase and use such processes within organisations should be trained **to "keep a handle" and a critical eye on algorithms**.

## Support research to develop studies to measure and methods to prevent bias

Available studies and analyses have started to show the magnitude of algorithm's discriminatory bias, but these still widely relate to systems implemented in the United States. In order to take account of the significantly more regulated and limited deployment of algorithms in Europe and of our specific demographic and social contexts, these analyses must be developed in the European Union and in France.

Public research organisations and public procurement could support these approaches and experiments which require real statistical expertise and a cross-disciplinary approach combining computer engineering (to understand and handle issues), economy (to measure any potential discrimination) and the law (to qualify the discrimination).

Exploring "fair learning" perspectives, i.e. the design of algorithms meeting equality and explainability objectives and not merely performance objectives, is another major research challenge.

[25] These guidelines are interdependent and reinforce one another: *Diversity, non-discrimination and fairness (no. 5), Accountability and Transparency (n°7), human agency and oversight*, guideline (no. 1) for example.

[26] Following the deployment of the GDPR, in February 2020, the European Commission published a white paper promoting an AI approach based on trust, the recommendations in which are largely drawn from works of a European expert group.

[27] Sarah Myers West, Meredith Whittaker and Kate Crawford, *Discriminating systems: Gender, Race and Power in AI*, AI Now, New York University, April 2019.

# Reinforce algorithms' information, transparency and explainability requirements

The user's right to information on the one hand, and transparency and explainability on the other, are clear pre-conditions to measure discrimination, monitor systems and ensure the effectiveness of the right to recourse. However, the opacity of systems and their secret nature are an obstacle to the discovery of potential biases and to recourse; this obstacle is all the more troublesome when the algorithm's result condition access to fundamental rights and public services.

The GDPR provides the first substantial solutions to these issues. For example, for reasons of transparency, its Article 13 sets out the obligation of providing "meaningful information about the logic involved" in any automated decision-making with a significant impact on the data subject. Furthermore the CRPA (Code on Relations between the Public and the Administration) - completed by the Digital Republic Act of 2018, specifies which information must be provided to the recipient of the individual decision regarding the "*degree and method of contribution of the algorithmic processing to the decision-making process*", "*the data processed and its source*", and "*the processing parameters and [...] weighting applied to the data subject*"[28].

To combat discriminatory bias, the legal requirements of information, transparency and explainability should be further developed.

Firstly, all of these requirements should not only be restricted to decision-making algorithms and those involving personal data processing[29]. Furthermore, they should be applied to both private and public sector algorithms. Lastly, the different requirements based on the level of automation of decisions should be reviewed: human intervention - which is sometimes formally provided for in many algorithmic processing operations - should not be merely symbolic and only actually provide artificial protection.

When they exist, transparency requirements in respect of third parties are still insufficient as noted by the Constitutional Council in its decision of 3 April 2020 on Parcoursup[30]. Third parties and not only the recipients of individual decisions should be able to access the criteria used by the algorithm to allow them to detect potential cases of bias.

The general information published on algorithmic processing and the individual explanations regarding a given decision must, in all cases, be provided to the public and to users in an accessible and intelligible language.

Professionals involved in algorithmic processes - whether employees or public servants - must be informed so that they are able to understand the tool's general operation, increase their vigilance as regards the risk of bias and ensure that they have effective control over the processing.

[28] Article R. 311-3-1-2 of the Code on relations between the public and the administration

[29] Collective report ordered by the Etalab mission, Ethique et responsabilité des algorithmes publics, ENA, Class of 2018-2019 "Molière", June 2019.

[30] Decision no. 2020-834 Priority preliminary ruling on the issue of constitutionality of 3 April 2020. The judges considered that the limits imposed by law to the exercise of the right to access administrative documents were justified by legitimate interest grounds and proportionate to this objective, i.e. the secrecy of deliberations protecting the independence of educational teams and the authority of their decisions. However, it makes one important reservation: each higher education establishment "*must account for - to use the terms of Article 15 of the declaration - of the criteria that it has used, where applicable using algorithmic means of processing, to study the applications sent on Parcoursup.*"(French Constitutional Council commentary, p. 26). See also Defender of Rights, Decision 2019-099 of 8 April 2019 on the operation of the Parcoursup platform, particularly the lack of transparency of the allocation procedure and the rejection of the request for communication of the algorithmic procedures used by the association made by the French Conseil d'Etat (12 June 2019, no. 427916).

# Perform impact assessments to anticipate algorithms' discriminatory effects

The principle of explainability and the identification of potential bias seem to clash with the "black boxes" that many algorithms become when the secret regarding the code is not revealed or when a learning algorithm is opaque. The issue of monitoring the effects of these systems must therefore be resolved from the algorithm's design phase or during their learning phase.

In Canada, audits including discrimination issues are required of public institutions since 1 April 2020 and the Federal Government has set up a platform - the AIA (algorithmic impact assessment) - to assist administrations with these impact assessments[31]. Such a requirement could be introduced in France based on the Data protection impact assessment (DPIA) model already provided for by Article 35 of the GDPR. This prior analysis, which is mandatory for some algorithms, must include an assessment of risks to the rights and freedoms of individuals and is therefore already a means of anticipating such discriminatory effects. However, expressly providing for the assessment of these biases as part of impact assessments or rendering these mandatory for all algorithmic processing operations[32] would ensure effective compliance with the principle of non-discrimination.

In addition to prior assessment, the regular monitoring of algorithms' effects after deployment should be required based on the model applied to monitor the side-effects of medicines.

Many questions still remain and some answers must be clarified. In any case, the methods and means to be implemented must ensure respect for our fundamental rights and freedoms in the face of the technological and economic frenzy surrounding what is commonly referred to as "artificial intelligence". For example, should we - as suggested by the European Commission - adopt a risk-based approach, which would increase the level of requirement and monitoring based on an algorithm's use and expected impact[33]? Are audit or accreditation procedures enough to ensure our rights are respected? How can we prove discrimination? Should we set up - as recommended by the Council of Europe - an institutional or regulatory framework and algorithm standards according to main sectors[34]?

The Defender of Rights will continue its reflection on this topic and will contribute to the reflection carried out by public decision-makers, notably in partnership with the CNIL, but also with Etalab, CNNum, the CNCDH, academics having participated in the seminar and the European network Equinet. In this perspective, guaranteeing respect for the rights of every individual, and in particular the right to not be discriminated against, will be its only compass.

[31] AIA is part of a broader framework of the Act on public finance management and its Directive on automated decision-making, having entered into force on 26 November 2018. This test sets out the responsibilities of federal institutions as regards the use of automated decision-making systems for administrative decisions.

[32] See on the CNIL's website, the infographic on algorithms for which DPIAs are required which repeats the positions adopted by the European Data Protection Committee (WP29).

[33] White Paper on Artificial Intelligence, *A European approach to excellence and trust*, European Commission, COM (2020) 65 final.

[34] Reco 1.4 (B) Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems.

Defender of Rights

TSA 90716 - 75334 Paris Cedex 07

Tel.: 09 69 39 00 00

www.defenseurdesdroits.fr

Find all our news at:

www.defenseurdesdroits.fr

Défenseurdesdroits

RÉPUBLIQUE FRANÇAISE