European Journal of Law and Technology, Vol 9, Issue 3, 2018

EJLT European Journal of Law and Technology

Algorithmic Decision Systems in the Health and Justice Sectors: Certification and Explanations for Algorithms in European and French Law

Sonia Desmoulin-Canselier[1], Daniel Le Métayer[2]

Abstract

Algorithmic decision systems are already used in many everyday tools and services on the Internet. They also play an increasing role in many situations in which people's lives and rights are strongly affected, such as job and loans applications, but also medical diagnosis and therapeutic choices, or legal advice and court decisions. This evolution gives rise to a whole range of questions. In this paper, we argue that certification and explanation are two complementary means of strengthening the European legal framework and enhancing trust in algorithmic decision systems. The former can be seen as the delegation of the task of checking certain criteria to an authority, while the latter allows the stakeholders themselves (for example, developers, users and decision-subjects) to understand the results or the logic of the system. We explore potential legal requirements of accountability in this sense and their effective implementation. These two aspects are tackled from the perspective of the European and French legal frameworks. We focus on two particularly sensitive application domains, namely the medical and legal sectors.

Keywords

Algorithm, algorithmic decision system, accountability, intelligibility, certification, medical device, explanation, machine learning, justice, health

1. Introduction

The tremendous development of data collection combined with the considerable progress in the field of artificial intelligence has led to a variety of data processing techniques that have enabled the partial or total automation of decision-making processes. Classification algorithms are already at work in many everyday tools and services on the internet, including search engines, social networks, comparison websites and targeted advertising. In the near future, they will play an even greater role in many situations in which people's lives and rights are strongly affected, such as job applications, loan applications, medical diagnosis, therapeutic choices, legal advice and, maybe, court decisions. This evolution gives rise to a whole range of questions. How relevant are the decisions taken on the basis of algorithmic processing? What are the risks of discrimination, of loss of autonomy for individuals or violation of fundamental rights and freedoms? What are the respective responsibilities of the developers of the algorithms and their users? The importance of these issues calls for an in-depth reflection on the technical solutions and the applicable legal frameworks.

'*Algorithmic transparency'* is often put forward as a necessary step to address the issues raised by algorithmic decision systems. However, this expression is open to varying interpretations. Generally speaking, it conveys the idea that algorithmic processing should be made more scrutable and understandable. This should mean not only informing people about the operations that have been or could be carried out but also explaining and justifying them. This is of a great importance, because it is a key condition to make the best use of algorithmic decision systems.[3] On one hand, this will help their designers and developers to improve them and ensure that they meet expected quality criteria. On the other hand, it will make it possible for non-expert users and people affected by the decisions (jobseekers, patients, defendants, etc.) to challenge their results. However, the polysemic notion of 'transparency' can lead to some misunderstandings. For example, it can give the impression that a complete description of the process is required, which would be incompatible with any form of industrial (or commercial) secret. Of course, such a demand does not seem realistic. Because the notion of 'transparency' is too imprecise, it can lead to ambiguities and be open to criticism without necessarily addressing the actual needs (Ananny and Crawford 2016).[4] Likewise, the concept of 'fairness' refers both to the idea that the software should correctly provide the expected functionalities (contractual expectations) and should comply with legal, moral and ethical codes (including fundamental rights and freedoms). The EU Data Protection Regulation (2016) states that '*Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject ("lawfulness, fairness and transparency*")', [5] which is consistent with national laws like the French 'Loi Informatique et Liberté'.[6] This wording seems to draw a distinction between 'lawfulness' and 'fairness' in data processing, but it is not easy to determine whether the latter term only refers to compliance with expectations or whether it also includes compliance with certain moral norms. Even though these terms are used in data protection legal texts and appear in various official documents[7], it is therefore useful to search for more suitable concepts.

A benefit of the concepts of 'intelligibility' and 'explainability' is to refer implicitly to requirements of accessibility and clarity. However, they do not render the idea of 'fairness' nor do they highlight the links with legal and moral principles. For this reason, the notion of 'accountability', albeit already used with specific definitions in accountancy and business law, seems more appropriate (Diakopoulos 2016, Binns 2017).[8] Reuben Binns defines accountability as follows: '*a party A is accountable to a party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate*'. Binns notes that '*in the context of algorithmic decision-making, an accountable decision-maker must provide its decision-subjects with reasons and explanations for the design and operation of its automated decision-making system*'. However, this application of the concept to decision-making focuses on two parties only, the decision-maker and the decision-subject. It does not consider other key stakeholders: the designer of the algorithmic decision system (who is generally different from the decision-maker), the users of the algorithmic decision system (who may be different from the decision-subject, for example, when the system is used by professionals such as practitioners or judges) and potential trusted third parties such as certification authorities.

In fact, the first party who should be accountable for an algorithmic decision system is its designer and the justification C in Binn's definition can be provided to different parties: to the decision-subject himself, to the user or operator of the algorithmic decision system or to intermediate trusted third parties such as certification bodies. From a legal point of view, accountability could therefore take at least two different but complementary forms. The first one is *certification*, defined as the obligation to demonstrate, for example to a certification authority or an auditor, that the algorithmic decision system meets specific criteria. These criteria can typically include accuracy (relevance of the results), absence of discrimination or other forms of fairness. The second one is a *requirement for explanations*, defined as the ability to ensure some form of intelligibility regarding algorithmic decision systems and to explain their results. This requirement could be met in different ways depending on the target audience. For individuals without technical expertise, this may involve the logical justification for particular results that are relevant to them. An expert, on the other hand, may also be interested in more global measures, such as explanations in the form of decision trees or other graphical representations showing the criteria taken into account by the algorithm and their influence on the results. Therefore, producing an explanation does not necessarily mean publishing the text of an algorithm or the source code of a piece of software, which can be impenetrable for the average person (and even for the experts). Furthermore, the results of algorithmic decision systems relying on machine learning cannot be understood independently of their training data sets. Indeed, these data can reflect biases that will be 'learned' and then reproduced by the algorithm.

Certification and explanation are thus complementary means of enhancing trust in algorithmic decision systems. The former can be seen as the delegation of the task of checking certain criteria to an authority (which is assumed to be trustworthy), while the latter allows the stakeholders themselves (for example, the developer, user or decision-subject) to understand the results or the logic of the system. In this article, we explore potential legal requirements of accountability in this sense and their effective implementation. These two aspects are tackled from the perspective of the European and French legal frameworks. We focus on two particularly sensitive application domains, namely the medical and legal fields. The rationale for the choice of these two areas is discussed in the following section. Section 3 reviews the legal and technical resources for the certification of algorithmic decision systems. Section 4 explores legal and technical solutions regarding the explanation of algorithmic decision systems. A conclusion synthesises our observations and suggestions, with some additional remarks.

2. Automated decision making and decision support algorithms in medical and legal matters: hopes, fears and the demand for accountability

Many reasons motivate our decision to use the medical and legal domains as case studies. Foremost among them was the observation that there has been an increase over the past few years in the supply of algorithmic decision systems and data processing tools in these sectors. These are areas in which the human stakes are particularly high since the decisions taken by medical doctors and judges can significantly affect the lives of the people concerned. Such sensitivities explain both the expectations people have of algorithms (particularly in terms of improving the soundness of decisions) and the fears that they can give rise to. Algorithmic decision systems currently available are very varied, ranging from those that optimise information for the decision maker (for example, by speeding up data retrieval and/or increasing the amount of data processed) to those that make autonomous decisions by running a computer program (supervised or unsupervised). They coexist in all their varieties, and, contrary to received wisdom, the latest innovations in artificial intelligence are not systematically replacing the 'expert systems'-type of computer programs. It is not always easy to distinguish between what are just slight amendments to pre-existing tools – like, for example, the use of a search engine on a jurisprudence database or the plotting of a decision tree that has already been validated by a national health authority –, and truly innovative propositions – like, for example, artificial intelligence software, such as IBM Watson Health for medical diagnoses.[9]

Healthcare and justice are traditionally said to be arts since they require technical skills and know-how-to-be wisdom as well as scientific knowledge. As Ricoeur said, they both involve applying a complex set of expertise and knowledge to a specific domain in order to arrive at a decision (Ricoeur 2001, p. 251-253).[10] In both domains, the appeal of objective proof is very strong because the risk of error is high and the stakes can be dramatic. An error is more difficult to accept nowadays when so much information seems to be available. However, an error is also more likely to occur now as a result of the huge, changing and sometimes contradictory mass of data and information that has to be processed in these times of economic and budgetary constraints. It is legitimate to try to make available as much information as possible so that physicians and other healthcare professionals, as well as judges and lawyers, can choose the most appropriate and therefore the most legitimate way forward, and any technological proposition in this sense is welcomed. The importance conferred on data in the two domains (state-of-the-art medicine, on the one hand, and the rule of statutory and case law, on the other) to justify any decision, explains why expert system development projects have been on the rise over the last four decades to support medical[11] and judicial decision.[12] The desire to reduce subjectivity in order to limit the risks of irrational decisional biases and to address territorial inequalities (between different jurisdictions or hospital trusts) also serves to justify a systematised use of algorithmic tools, which act as standardisers. However, this second argument can lead to fears that, in the long run, the algorithmic rationality would totally replace, at some point, human decisions. This then raises the question of how a human decision can be justified when it is based on a data processing tool that is too complex for the user to be able to give an account of the reasons underpinning that decision. In other words, in domains as sensitive as health and law, can we accept the use of tools whose results, no matter how precise, the users are not able to explain?

These preliminary remarks highlight two crucial distinctions. On the one hand, there is a difference between 'decision' and 'decision support' or 'decision aid'. On the other, there is a distinction between the professional user and the decision-subject. First of all, most of the algorithmic decision systems currently on the market for legal and medical usages are presented as 'decision support systems'. While some of these algorithmic devices can function as autonomous decisional systems, they are only currently being offered as an assistance or support aid. This is an important point to highlight because the use of an automatic decisional algorithmic system in legal and medical matters may sometimes conflict with European and French regulations. Only members of the medical profession are legally authorised to practise medicine[13], and any person '*involved habitually […], even in the presence of a doctor, in establishing a diagnosis or treating illnesses, congenital or acquired, real or perceived, through personal acts or verbal or written consultations or through any other means, whatever they may be*' is liable to prosecution in France (as stated by the French Public Health Code).[14] A non-restrictive interpretation of this statement has been adopted in case law, and various kind of acts are covered by the prohibition of the illegal practice of medicine.[15] Medical doctors assume, either individually or in conjunction with the establishment that employs them, responsibility for their own diagnoses and therapeutic choices. Ethical and legal norms converge here. They are supposed to make informed decisions in conscience and to justify them in cases of damage or misconduct.[16] As stated by the French Code of Medical Ethics, they '*shall not alienate their professional independence in any way whatsoever*'.[17] Automated decision-making systems which are not functioning as medical devices implemented by doctors are therefore exposed to illegality. Judicial independence, embodied in the judge's decision taken in conscience, is imposed with the same force. Only magistrates are entitled to serve justice. In France, several national legal sources may be invoked here: from procedural code[18] to data protection law.[19] Individual decisions producing legal effects, which are obtained solely from the results of algorithmic processing, are in principle contrary to the EU and French data protection laws, with exceptions. The EU Directive 95/46/EC of 24 October 1995[20] and its replacement, Regulation 2016/679 of 27 April 2016 (General Data Protection Regulation: GDPR)[21], specifically state that natural persons have the right not to be subjected to, or be the subject of, a decision based solely on automated processing (including profiling) where this automated processing either produces '*legal effects*' concerning those persons or '*significantly affects him or her*'. The exception to this is when the decision has been adopted in the performance of a contract or within a legal framework stating that there are '*suitable measures to safeguard his [or her] legitimate interests*' (1995 version) or '*appropriate measures to protect the data subject's rights and freedoms and legitimate interests*' (2016 version, applicable from 25th May 2018). Outside of any contractual relationship, the data subject's consent may also authorise the use of automated decision making, but with the following option of a similar condition attached: '*the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision*'. While this leaves states some margin for interpretation in regard to organising '*suitable measures*' for safeguarding individual rights and freedoms, which determine the validity of a law authorising such automated processing, the regulatory terms clearly indicate the need for a special law. In the absence of any specific legal framework then the use of a decisional algorithm in judicial matters would be contrary to EU regulation. French law takes much the same approach. Article 10 of the Loi Informatique et Libertés (French Data Protection Act 1978, amended June 2018) prohibits the use of automated decision making – i.e. decisional algorithms – in judicial matters where that decision involves '*an assessment of a person's behaviour*' and where the '*automated processing of personal data*' is '*aimed at evaluating certain aspects of their personality*'. For other decisions that '*produce legal effects concerning the natural person*', it is expected that they cannot be taken '*solely on the basis of the automated processing of data, including profiling*'. That excludes the use of decisional algorithms for some litigations (but not all, because some legal matters do not involve '*behaviour assessment*' and some algorithms do not deal with the '*profile of the data subject*'). However, it is important to notice that these stipulations do not apply to decision support algorithmic systems since these systems do not automatically produce a decision but rather formulate recommendations. This exclusion is sometimes justified by the idea that the important thing is that human intervention has taken place. Article 22 GDPR is worded along these lines when requiring that data subjects should be granted with the right to obtain post-decision human intervention (for contestation purposes). Nevertheless, one may ask if such a focus solely on automated processing is really justified. It could lead to neglect the importance of the explanation and certification issues for decision-aid algorithms. As already mentioned, most algorithmic processing currently available in the medical and legal fields promises to act as a decision support tool. But, it can only truly be a valuable support if the user understands what the instrument does. It is therefore imperative to ask what explanations should be provided and what verifications should be carried out for decision support systems. An examination of the legal resources and technological solutions relating to the need for verification and explanation (i.e. accountability) of these algorithms is therefore not just useful, it is essential.

The second crucial point to raise here concerns those on the receiving end of this demand for accountability. As suggested by the textual references already mentioned, most legal provisions dealing with the information and explanations that should be provided come from personal data protection laws. These give natural persons whose personal data have been collected and processed (which makes their identification possible, either directly or indirectly) the right to be informed on their use and to demand that rectifications are made and that certain uses are prohibited. The limited scope for applying these regulations means they are of limited use in terms of what we are focusing on here. On the one hand, they do not apply to algorithmic systems that use anonymised data. Hence, it could be argued that, for example, in judicial matter a decision support algorithm dealing only with anonymised legal decisions would not be regulated. On the other, the *professional users* are not entitled to exercise the rights conferred by this set of regulations. Data protection laws are not suitable resources to justify imposing an obligation of intelligibility or non-bias result verification in favour of the users on the developers and providers of medical and legal decision support systems. Yet, it is important for the patient, the litigant or the respondent to know that the physician or the judge understands the results produced by the algorithm (or at the very least, detects its limitations) and that they can check the potential biases in its functioning. In the next two sections, we successively analyse two variations of this requirement for accountability, namely certification (meaning obligation to justify to the experts) and explanation (meaning requirement to justify to the users).

3. Towards a certification obligation for algorithmic decision systems

In our opinion, it would be useful if the verifications relating to both the proper functioning of an algorithmic decision system and its conformity to moral and legal principles (absence of discriminatory effect, fairness, and so on) could be carried out before the product is launched. There are already certification systems in place that determine market access for medical devices and which control the use of some judicial expertise tools. It is therefore pertinent to examine the legal resources and technological solutions that are currently available. Since legal regulations are organised on a sectoral basis, this section will examine certification requirements first in the medical (3.1) and then in the legal domain (3.2). The section will close with a discussion of some interesting technological solutions (3.3).

3.1. Certification requirements for medical devices and prescription management software

The different software made available[22] within the EU may be subject to EU laws covering medical devices (MD) when they are intended for medical use even if they were developed and manufactured outside of Europe. Any MD entering the European market must be certified. Its manufacturer must fulfil a number of essential requirements, and, in particular, they must set up, document, implement and maintain a 'risk management system'. Directive 93/42/EEC[23], which will continue to serve as a reference for certificates until 2020, 2022, 2024 or 2025 (as appropriate)[24], and its replacement, Regulation (EU) 2017/745 of 5 April 2017[25], retain a broad, functionalist definition of MD. The medical applications targeted include, most notably, the diagnosis, prevention, monitoring, treatment and alleviation of illnesses plus (since the introduction of Regulation 2017/745, for clarification purposes rather than to add something new) prediction and prognosis. These regulations are therefore of direct interest to us.[26] However, there are many different kinds of difficulties to overcome. Some concern the applicability of MD regulations, while others relate to the problem of establishing which rules are applicable because different categories of MDs are organised (Class I, Class IIa, Class IIb and Class III).[27] The claims made by manufacturers and importers (about the strategic and commercial choices they have made as well as the technological characteristics of the tool) play an important role. Since the aim is to confirm that medical algorithmic decision systems come under MD law, it should be noted that applying criteria of specificity (namely 'intended by the manufacturer to be used for human beings for' diagnostic and/or therapeutic purposes or 'specific medical purposes') or of action on the human body can lead to differing interpretations. Hence, in France the national body in charge of drugs and MDs (*Agence Nationale de Sécurité du Médicament* et des Produits de Santé: ANSM) holds that for a software to be classified as an MD, it must carry out other functions than just storing, archiving, compressing, communicating and/or searching for data. It must also bring the patient individual benefit. A diagnostic support system is therefore only considered to be an MD when '*new medical information*' is produced, such as a treatment path suggestion, a patient-specific outcome or an alert function. Examples of software currently classified as MDs include a tool for segmenting images and generating files to determine target points for radiotherapy treatment, software for collecting and sorting alarms in an intensive care unit and a tool for predicting the risk of melanoma.[28] The guidelines do not, however, answer all the questions. Under the aegis of the former directive, the ANSM has generally excluded software that produces recommendations in the form of a '*generic result for a group of patients*' or which proposes '*functionalities aimed at verifying the absence of contraindications or of unrecommended medicinal combinations through a database*'.[29] The reason for this exclusion is not clear. Likewise, the department within ANSM that deals with MDs acknowledges the fact that some '*prescription support system and dispensing support system are nevertheless not considered to be medical devices*'.[30] At the European level, the European Court of Justice recently confirmed that '*a software, of which at least one of the functions makes it possible to use patient-specific data for the purposes, inter alia, of detecting contraindications, drug interactions and excessive doses, is, in respect of that function, a medical device within the meaning of those provisions, even if that software does not act directly in or on the human body*' (ECJ Case C-329/16: *Snitem & Philips v. France*).[31] And, for the future, the new European Regulation 2017/745 explicitly states that digital decision support systems, like other software, may fall in the category of 'active medical devices'. For some of the algorithmic decision systems (but not all of them), it would therefore be conceivable to draw on MD regulation to require verification.

This brings us to the question of the rules that are applicable to medical algorithmic decision systems. Indeed, depending on the class of MD, the regulatory expectations are very different. They range from simply having to compile a technical file to conducting clinical trials and having the certification verified by an independent body. Since there are a variety of medical algorithmic decision systems, the rules that apply will depend on both their characteristics and their context of use.[32] Classification may also involve description as an accessory[33], which means that the rules conditioning the marketing of the principal MD also apply to the accessory. The person who declares, registers or requests certification, in other words, the legal entity that intends to launch the product on the European market (manufacturer, importer, etc.), gives the original description of the MD through the claims made and through analyses successively proposed on the software's functioning (principal MD or accessory MD, risk analysis, etc.). While, in principle, a medical algorithmic decision software falls, as a priority, into the active MD Class IIa, the rules actually applied to it may therefore be different. The national records of declarations held by national authorities are informative on this point, although the data available were collected under Directive 93/42/EEC.[34] For future reference, Regulation 2017/745 states that '*software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause: - death or an irreversible deterioration of a person's state of health, in which case it is in class III; or - a serious deterioration of a person's state of health or a surgical intervention, in which case it is classified as class IIb. Software intended to monitor physiological processes is classified as class IIa, except if it is intended for monitoring of vital physiological parameters, where the nature of variations of those parameters is such that it could result in immediate danger to the patient, in which case it is classified as class Iib*'.[35] Despite its precision, the text does not cut interpretation off, especially regarding the risk of 'irreversible deterioriation'. One might indeed consider that such a risk is always possible when medical diagnostic or prescription is at stake. If a classification 'error' occurs, it can be identified and rectified by the national authority in charge of receiving the declaration. If there is a significant dispute, Regulation 2017/745 indicates that the Commission may refer, on their own initiative or on request from a member state, to the new Medical Device Coordination Group. European and national judges will also be able to bring about a reclassification in the event of a dispute between competing manufacturers or damage caused to users.

EU certification requires that the manufacturers or those responsible for bringing the product to market (or put into service, Regulation 2017/745) declare that they will comply with a certain number of regulatory obligations ('declaration of conformity') and that an independent body ('notified body') has carried out the necessary checks on the manufacturing processes, products and control systems put in place by the manufacturer. The compliance criteria for certification have been drawn from two sources. On the one hand, EU legislation imposes a number of fundamental requirements with legal force.[36] On the other, proof that these requirements have been complied with can be demonstrated through the implementation of technical standards developed by the various recognised bodies (ISO, ECS, and national standardisation associations like AFNOR in France). This is very often the case when the regulatory documents (Directive 93/42 and Regulation 2017/745) allow proof of compliance with these technical standards to also mean presumption of proof of compliance with regulatory obligations.[37] As regard 'essential' or 'general' regulatory requirements, they provide that MDs must achieve the expected performances, that they must be 'safe and effective' and that they must 'not compromise' patient safety.[38] This includes commitments that the tool should function correctly.[39] Considering medical decision software, Regulation 2017/745 precises that the responsible party must be committed to ensuring compatibility and interoperability[40], reliability of devices with diagnostic and measuring functions[41], validation of performances[42] and security measures to prevent unauthorised access.[43] Some provisions concerning the reliability of MDs in terms of their stated characteristics and performances, including a consideration of the passage of time (technological obsolescence), could be linked with the question of the 'fairness' of algorithmic systems. However, most of the questions that we are interested in (discrimination bias, explanation and intelligibility) are not explicitly covered. Class IIa MDs are subject to rules imposing a self assessment of, and sometimes a notified body control, on conformity and quality management. The annexes of Regulation 2017/745 require that the manufacturer create documentary files containing: the product description, its development methods, the 'necessary explanations' for understanding its functioning and the results of the risk analysis carried out (this is the so-called 'technical file' section). But, these requirements only concern basic information and do not seem to cover explanation about the algorithmic logic used or verification that results are not biased. Since MD laws focus only on market fluidity and people safety, a creative interpretation would be necessary to transform these requirements into resources available in respect of the accountability of algorithms. A coherent, but demanding, understanding of the reliability and safety objectives should lead to the idea that checking the absence of discriminatory bias and the intelligibility of the algorithmic logic is needed.

Although promising, considering its scope of application, the European certification legal framework does not provide yet enough strong elements to impose the new kind of accountability we need. The same kind of promising but deceptive content may be found in national laws. Thus, it is urgent to improve the European and national laws regarding MDs in order to cover the verification and explanation issues for all algorithmic decision systems. However, the MD legal framework already exists, and therefore could be used to support the implementation of algorithmic accountability. The situation is very different in the judicial domain.

3.2. Certification requirements for judicial algorithmic decision systems?

Some international standards[44] and a few national legal statements aim at ensuring the reliability and quality of forensic work or judicial expert assessment, but their areas of application are very limited. For example, the processing of DNA material for evidential purposes within the judicial context is entrusted only to registered laboratories[45], which must respect technical standards (such as the ISO IEC/17025 standard, which has been translated into guidelines by the International Laboratory Accreditation Cooperation and the European Network of Forensic Science Institutes).[46] These legal regulations guaranteeing the reliability of results do not apply outside of their strict remit. Consequently, they do not concern judicial algorithmic decision systems. Hence, even though some technical standards might be relevant to our topic (for example, the *Standard Guide for Establishing Confidence in Digital Forensic Results by Error Mitigation Analysis*) [47], they are not compulsory in all European countries. In France, for instance, except when specific rules apply, it is left to the judge's discretion to determine the useful sources of information for uncovering the truth and for dealing with cases of dispute resolution. That is why, in administrative and civil matters (which includes compensation for the victims of criminal offences), judges can use damage classifications and compensation scoring tables. These compensation tables may be seen as very rudimentary forms of algorithms in that they formalise the procedure, on the one hand, for arriving at figures that are based on expert knowledge and, on the other, for rationalising and standardising court rulings.[48] As already mentioned, the personal data protection regulations proscribe, on principle, any recourse to automated court decisions except under precise circumstances (GDPR, Article 22, aforementioned). More generally, French jurisprudence affirms that a court ruling cannot be based solely on a compensation scoring table.[49] In administrative matters, some regulatory procedures integrate risk calculation algorithms, but they always reserve the judges' appraisals and the possibility of justifying full compensation based on other criteria than those retained by the calculation method. This is why the French Conseil d'État (Council of State: Supreme Court in administrative law matters) regularly hears appeals against Committee decisions on compensation for nuclear accident victims that are based on a method of calculating the dose of ionising radiation received, an algorithmic method of risk calculation that is used to allow plaintiffs to benefit from the presumption of a causal link between the illnesses and pain they complain of and the nuclear tests carried out by the French state. When the algorithmic result concludes that the risk is negligible, the plaintiff must assume the burden of proving the causal link. The Committee's decisions are subject to appeal before the administrative jurisdictions, and the Conseil d'Etat regularly points out that the administrative courts of appeal assess the elements submitted to them with sovereign power, including the results from the aforementioned algorithmic method of risk calculation, which constitutes just one of a number of elements.[50] Despite attempts in some lawyers' arguments to raise objections[51], the reliability of this calculation method has not been the subject of a proper discussion in the justice system. Its pertinence is left to the discretion of the Committee, and the judges hearing appeals are happy just to verify that the result is not the only element justifying the decision. The use of algorithmic decision systems is indeed admissable in judicial matters as long as their results are only one element in a body of evidence and source of information.

To our knowledge, there is no certification requirement applyied to algorithmic decision systems used for judicial purposes. For a long time, the fact that algorithms are used merely as 'aids' seemed to preclude the need to draw up guarantees of reliability, 'intelligibility' and 'fairness'. Such a position, however, appears highly questionable. As stated by Pr. Cadiet in a recent report, '*The algorithmic processing results should be submit to question and put in perspective. Yet, "prediction" tools, if they suggest a solution to those who use them, do not reveal and explain their logic and the method followed to reach the solution. […] A regulation is necessary. This regulation could, in a first view, take place through an obligation of "transparency" for algorithms. A control by public authorities could also be put in place, under reserve to be flexible enough. Finally, it could also be a quality certification by an independent body*' (Cadiet 2018, p. 25).[52] A difficulty here may arise from the current permissiveness regarding the various compensation scoring tables and damage classifications available. For the moment, their use is at the discretion of the judges, and, as such, they are neither compulsory, systematic nor standardised. Outside of specific procedures, such as the aforementioned compensation for the victims of ionising radiation, it seems that some of these tools are indirectly imposed through their standardised use by judicial experts in the medical field.[53] However, these are more the working practices of judicial experts than of judges. This leads us to draw a parallel here with the rules applicable to judicial expertise tasks, allowing us to move the reflection forward. In a way, the service provided by algorithmic decision support systems appears to be similar to expert assessment. As summarised by Vergès, Vial and Leclerc, '*expertise is the measure by which a person [the judge or a party] entrusts another (the expert) with the task of clarifying a technical question for them in order to help them make a decision that is encumbent upon them*' (Vergès, Vial and Leclerc 2015, p. 669).[54] Beyond the terminological analogy between 'judicial expert' and 'expert system' for decision support, there is clearly a similarity in the functions attributed to these two kinds of 'aids'. In France, on principle, judicial experts are chosen on a list (that is valid either nationally or within a jurisdiction).[55] Such is the case in civil and criminal[56] and administrative[57] matters. Entry in a list of experts is not, however, always a necessary precondition,[58] nor is it a systematic guarantee of competence. Consequently, this raises questions around the persistance of this setting if a certification were to be put in place for the 'expert systems' and other algorithmic decision tools. The idea of systematising the use of damage classifications and compensation scoring tables (particularly in the form of reference tables[59]), that has been suggested and mooted for some time,[60] generate the same kind of doubts.

At the moment, existing legal requirements regarding judicial expertise and compensation scoring tables do not give a strong support to claims that algorithmic decision support systems have to be certified (with stringent obligations). However, the algorithmic decision tools have a much broader spectrum of actions than compensation scoring tables and could work more effectively to address geographic disparities and judgment biases. They also may be seen as more objective, but less intelligible. Some French jurisdictions have already tried one of these decision systems, with unsatisfying results, perceived as difficult to understand.[61] The idea to use more widely these tools, even as non-compulsory aids, seems to gain more supporters, but there is still a need to enhance trust in algorithms. It is therefore necessary, as stressed by Cadiet, to start a reflection concerning their certification, the verification of their fairness and the need for explanation (see section 4). Even confined in an indicative role, algorithmic decision systems are highly technical tools and they might strip the judges of their essential critical capacity. It would be paradoxical to advocate the use of decision support systems for combatting judicial disparities without checking that those systems do not themselves contain biases (visible or invisible) or produce discriminations. Some technical solutions are conceivable in this respect.

3.3. Technical solutions to detect and limit biases and discriminatory effects

The certification of a system is the verification that it meets well-defined technical standards and possibly also legal requirements. As a matter of fact, any certification must be based on a technical reference framework, that is, a specific set of reference criteria. Therefore, the first question regarding the certification of algorithmic decision systems relates to the definition of the technical criteria (3.3.1). The second question is the actual implementation of the certification process (3.3.2). For example, very different solutions are possible depending on the possibility of getting access to the text of the algorithm, the ability to control its input data and whether it is possible to take actions only a posteriori (after the algorithmic system has been deployed) or also a priori (before or during its development).

3.3.1. Criteria of discrimination

Discriminatory treatments are illegal in many countries for certain types of activities, such as employment, rental housing and bank loans.[62] The fact that algorithmic decision systems can lead to discriminations has already been studied in many areas, in particular in the justice system.[63] However, to decide whether an algorithm is acceptable or not from this point of view, it is necessary to define precisely what is meant by discrimination. One of the difficulties in this respect is the need to take into account not only direct discriminations but also indirect discriminations. A discrimination is indirect when the system does not exploit directly prohibited factors (such as, for example, ethnic origin) but uses other information that is correlated with these factors (for example, an individual's home address). Moreover, an algorithm may offer guarantees of non-discrimination in relation to two factors considered independently (for example gender and ethnic origin) but still lead to discriminations on the two factors considered simultaneously. Many definitions of discrimination have been put forward in scientific literature and in legal texts. Some of them are based on the probabilities of granting or denying a benefit.[64] Generally, a distinction is made between a protected group *G1* (for example, the black or female populations) and a non-protected group *G2* (for example, the white or male populations), and probabilities are compared in different ways. For example, let *P1* and *P2* be defined as follows:

- *P1*: probability of rejection for the protected group *G1*
- *P2*: probability of rejection for the non-protected group *G2*

Possible measures of discrimination are, for example:

- The difference of risks: *P1 - P2*
- The risk ratio: *P1 / P2*
- The opportunity ratio: *(1 – P1) / (1 – P2)*

As an illustration, the US federal law on employment requires that the opportunity ratio (also called 'disparate impact') should be greater than *80%*. This means that the opportunity for a person of the protected group to get the benefit must not be lower than *80%* of the opportunity for a person of a non-protected group. In UK law, the prohibition of discrimination on gender and ethnic origin is defined in terms of the difference of risks. These definitions, which are justifiable from a statistical point of view, do not, however, guarantee that 'equals are treated as equals', that is, that similar profiles are treated in a similar manner. One way of expressing this objective is to introduce notions of distances between the profiles and between the results of the algorithm. If we call these distances d and D respectively, then algorithm A applied to profiles x and y is considered non-discriminatory if $D(A(x),A(y)) \leq d(x,y)$, meaning that the distance between the results of the algorithm is limited by the distance between the profiles. Equal profiles (such as $d(x,y)=0$) will therefore be treated equally ($D(A(x),A(y))=0$, meaning that $A(x)$ and $A(y)$ are similar). Other definitions of discrimination take into account the fact that specific factors could be considered legitimate to justify differences of treatment (for example, a job that requires physical strength). Formal comparisons between certain measures have also been established. For example, it is possible to show that, in certain conditions, the above definition based on distances is stronger than 'disparate impact'.[65]

3.3.2. Prevention and detection of discrimination

Several research groups have worked on the introduction of non-discrimination requirements within the development phase of algorithmic decision systems. Different strategies are possible to achieve this goal, which focus on different stages of the process.[66] The first (upstream) option consists of filtering out any discriminatory bias from the training data sets. The second option consists of adapting the algorithm itself[67] to ensure that it does not produce discriminatory results. The third (downstream) option consists of correcting the results of the algorithmic decision system to avoid any potential biases. While the second option provides the most precise results[68], it is specific to each algorithm. The other two are more generic but often less precise.

When it is not possible to be proactive, that is, to take actions during the development of the algorithmic system, the only solution is to try to verify a posteriori that the algorithm does not to lead to any discriminatory treatments. The main problem with this approach is that the text of the algorithm, as well as the code of the program implementing it, are often protected by copyright law or industrial secret. Their developers or the stakeholders using them can therefore object to their disclosure. They can have different motivations to do so. For example, a company may consider that its algorithms represent a strategic business differentiator or may fear that users could exploit this information to their advantage by manipulating the algorithm.[69] The only option in this case is to carry out a so-called 'black box' analysis, that is, to experiment with the algorithm by providing inputs (when possible) and observing its outputs. The experimentation protocol must be precisely defined in order to guarantee the statistical value of the results. This approach has been adopted by a number of research groups to uncover the criteria used by targeted advertising systems. For example, the AdFisher[70] system allows researchers to simulate the behaviour of internet users with varying profiles, to observe the advertisements that they receive and to analyse the differences between them. This research has made it possible to detect a form of gender discrimination, where female profiles are less likely to receive advertisments for well-paid jobs. AdFisher also pointed out the limitations of the information provided by Google about the factors used to personalise its advertisements[71], which was clearly incomplete. Similarly, the Sunlight system[72] has allowed researchers to show that, contrary to its declarations, Google uses sensitive information (for example, health-related data) to target its advertisements.

The technical solutions described in this section show that the certification of algorithmic decision systems with respect to precise measures of 'fairness' is not out of reach, even if much progress has still to be made in this area. Certification can indirectly strengthen the trust that users and decision-subjects can have in these systems. However, both users and decision-subjects also have the legitimate right to require understandable information or explanations about the logic of these systems and justification of their results, especially when important decisions are at stake.

4. Towards a duty of explanation for the users of algorithmic decision systems

The certification should provide certain guarantees on the functioning of the algorithmic decision systems (for example, quality of results, absence of bias), but their users (such as a physician or a judge) should themselves be able to interpret the results and their limitations. This is a *sine qua non* of a truly useful use in such sensitive contexts as health and law. Moreover, the persons potentially affected by these results – the decision-subjects (patients, litigants and respondents) – are also entitled to expect basic information on the logic and criteria used. Accountability of algorithmic decision systems claims not only for verification, but also for explanation. Intelligibility is a key matter. Following the same intellectual pathway as in section 3 for certification, we will therefore investigate the legal resources that could potentially justify such a duty of explanation (4.1), before discussing the technical solutions and difficulties (4.2).

4.1. Legal requirements as regards explanations

The personal data protection legislation provides some textual basis for an obligation to give explanations to decision-subjects when algorithmic decision systems are used with legal (or similar) consequences. European Directive 95/46/EC and its replacement, the GDPR, enforce rights of information and access (as well as rectification), which include precisions relating to automated decision-making. Article 13, 2., (f) of the GDPR states that the data controller should provide the data subject, from which personal data are obtained, with the additional information '*necessary to ensure fair and transparent processing*', which includes '*the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*'. There is an equivalent provision covering personal data collected from a third party[73] and the right of the data subject to access their personal data.[74] In French national law, the Loi Informatique et Libertés includes similar requirements. However, as it has been emphasised in recent publications, data sujects are not entitled to a full explanation of the individual decision affecting them, but only with a right of information and access (Wachter S., Mittelstadt B. and Floridi L. 2017; Goodman B. and Flaxman S. 2016).[75] It merely includes general explanations on the global logic of the algorithmic decision system functioning. Furthermore, these provisions only concern automated decision-making processing personal data. They only aim at protecting data subjects (from a wrongful use of their personal data), not at protecting all litigants and respondents from unintelligible decisions. They do not cover algorithms that use anonymised data. They do not apply to decision *support* systems, which do not automatically produce decisions. Some authors argue in favour of a more generous interpretation covering all kind of algorithmic decision software, but their position seems fragile (Malgieri and Comand 2017).[76] The last important element to notice is that GDPR provisions are not dedicated to the users of algorithmic decision system. Therefore, they have no impact on the intelligibility of algorithmic decision support systems for medical doctors or judges who may use them. On this issue, the European legal framework seems at the moment incomplete or deficient.

May national legislation supplement this blind spot? In France, the Law for a Digital Republic (Act No 2016-1321 of 7 October 2016) and its *décret d'application* (implementation order) [77] enforce the inclusion of an '*explicit mention of the use of algorithmic processing within the context of an administrative decision'* and the '*possibility for the user to claim for an explanation on its functioning*'. These provisions are interesting, first, because they are aimed at decisions taken '*on the basis of algorithmic processing*', which can be interpreted more broadly than '*automated decision-making*', and, second, because this preoccupation with explaining algorithmic processing emerges outside from the context of privacy and data protection law. Thus, it stresses out the transversality of the issues at stake, which are not only about protecting data subjects against a wrongful use of their personal data. Articles R. 311-3-1-1 et seq. of the French code on the relations between the administration and the public give an idea of the type of regulatory expectation that can be formulated. Article R. 311-3-1-2 requires that '*at the request of the person who is the subject of an individual decision made on the basis of algorithmic processing, the administration communicates to them, in an intelligible form and without infringing any secrets protected by law, the following information: 1) The extent and mode of the contribution made by the algorithmic processing to the decision-making process; 2) The data used and their sources; 3) The processing parameters and, where applicable, their weighting applied to the data subject's situation; 4) The operations performed by the processing*'. Depending on the technical approach followed and on the interpretation adopted, it seems that these requirements can either be easily met or may pose an insurmountable technical conundrum. However, these regulations only shed indirect light on our topic, because they only concern decision-subject information about administrative individual decision. They do not constitute the foundations on which to build a duty of explanation for the benefit of the users of algorithmic decision systems and the decision-subjects in medical and judicial matters.

Should we conclude, then, that there is an important breach in our legal framework regarding the intelligibility of algorithmic decision systems, especially for their users? If no legal foundation may be found for such an obligation, it would be an astonishing conclusion given the importance of ensuring an intelligent and critical use of these new tools. Many recommendations, from various origins (ethics committees like CERNA in France[78]; French authority in charge of Privacy and Data protection: CNIL[79]; European Parliament[80]), have highlighted how important it is that algorithmic data processing and artificial intelligence do not lead to decisions that are incomprehensible to human intelligence. At the moment, it seems that legal requirements on this matter may only be found in general principles regulating medical and judicial practices in national and international laws. In regard to medical matters, the legal and ethical principle of responsibility is based on a decision made 'in conscience'. Accordingly, a physician shall be given the opportunity to justify his decision with the information necessary to critically consider the different therapeutic options suggested by the algorithmic system. The existing legal framework could be strengthened by clarifying and modifying MD regulation in order to introduce clear and strong provisions about explanations that should be given on the functioning of the algorithmic decision device, as it has been already mentioned for verification. MD regulation should compel the producer or manufacturer to provide the users information about the algorithm's logic and the kind of data processed. With regards to judicial decision-making, several legal principles may serve as guides, most notably the motivation and adversarial principles. As a reminder, the adversarial principle has also been imposed in judicial expertise matters since the European Court of Human Rights recognised the preponderant role of expert conclusions in court rulings.[81] Litigants and respondents should be made aware of the kind of tools that are used to uncover the facts or the jurisprudence, and they should be in position to discuss their influence on the outcome of the litigation. In terms of the need for reasoned judgments made in conscience, the decision shall be taken only by the judge, and the reasons (of fact and of law) justifying the decision shall be exposed. The motivation principle is an essential element of the judicial process and a basic guarantee for the litigant or the respondent, who needs to understand the decision to accept it or to contest it.[82] Mirroring the '*constitutional aim of ensuring the law is accessible and intelligible*' (acknowledged by the French Conseil Constitutionnel (Constitutional Court))[83], the obligation to motivate and the need to explain the decision applies to all courts (from the lower to the supreme court – the Cour de Cassation in France).[84] Thus, algorithmic decision systems used in courts should not counteract this command by clouding the decision-making process.

There is no reason to consider that the general principles regarding medical and judicial practices do not cover algorithmic decision tools. How could medical and judicial decisions be taken in conscience and how could they be reasoned and subject to responsibility (at least in disciplinary terms) if the professionals concerned are satisified with using results or recommendations that they do not even know the global logic of? However, these legal requirements remain general principles and, in France and Europe, no legal statement explicitly refers to explanations on the functioning of algorithmic decision systems that should be given to the users. In order to enhance trust in this matter, it could therefore be appropriate to adopt new specific provisions. The need for explanation could be part of the verification process, described in section 3, but it could also be useful to add legal provisions dedicated to algorithmic decision systems in medical and procedural laws. This leads us to the discussion of the technical approaches available for enlightening users and data subjects on the functioning of the algorithmic system.

4.2. Technical solutions and challenges

The implementation of legal obligations to explain algorithms raises a number of challenges, in particular on the technical side. Some systems are based on techniques like deep learning, the results of which are intrinsically difficult to understand. Moreover, the internal model built by a system may constantly evolve when learning continues during the exploitation phase. Any explanation of this kind of system would therefore be valid only for a particular point in time. In addition, some algorithms or their parameters are also regularly amended or adjusted by their users, which further complicates the explanation task. Finally, the large quantity of factors taken into account by a system can be an additional obstacle to intelligibility. Generally speaking, certification and explanation can be viewed as different ways to answer questions about the algorithms. If they are precise and accurate enough, these answers may provide more control over algorithmic decision systems. Different types of questions can be relevant depending on the situation. For the sake of simplicity, we have grouped these questions into two broad categories, namely questions on the global logic underpinning the algorithmic system (4.2.1) and more local questions concerning particular cases (4.2.2). We then compare these two options (4.2.3) before analysing the ways in which these explanations can be implemented (4.2.4).

4.2.1. Explanations on the global logic of an algorithm

A way of improving the intelligibility of an algorithm is to provide a description of its logic, or, more precisely, of the logic of the underlying model. This is a right that is explicitly provided for in both the Loi Informatique et Libertés[85] and the GDPR[86], although the effectiveness of this right is debated.[87] The logic of a model can be described in different ways. The most common is a graphical representation in the form of decision trees. As an example, Figure 1 represents a hypothetical decision tree concerning parole court decisions. The nodes of the tree represent criteria or questions (for example, 'is the prisoner a recidivist?'), which are posed in turn (in a top-down fashion). The leaves of the tree represent decisions (for example, 'rejection') or questions leading to other decision trees (for example, 'serious efforts at social rehabilitation?').

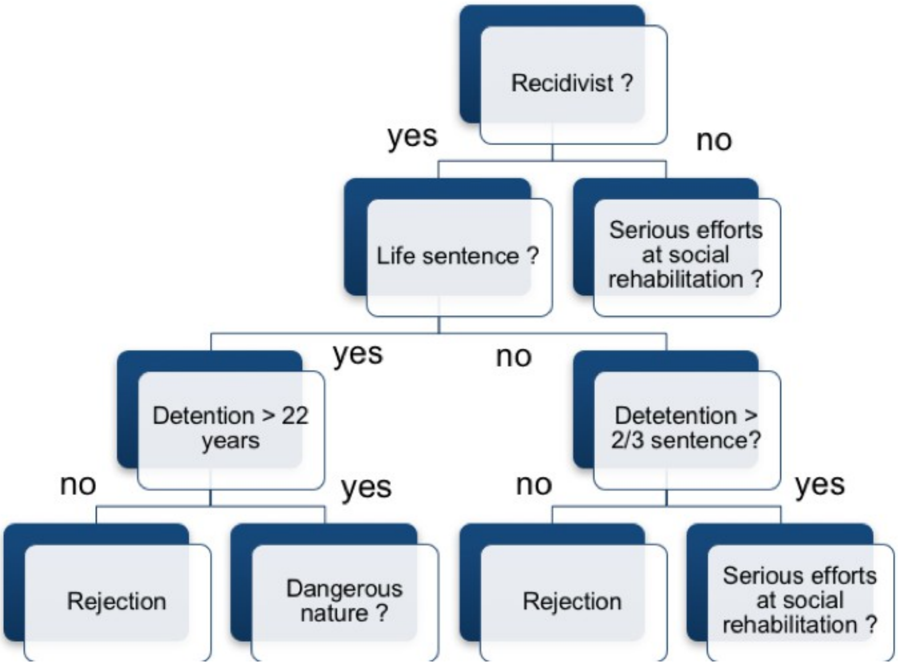

Recidivist ?
yes
no
Life sentence ?
Serious efforts at social rehabilitation ?
yes
no
Detention > 22 years
Detetention > 2/3 sentence?
no
yes
no
yes
Rejection
Dangerous nature ?
Rejection
Serious efforts at social rehabilitation ?

Figure 1

A decision tree can also be transformed into a set of of rules in natural language (for example, ‘If the prisoner is a recidivist, then one of two cases will apply. Either he was sentenced to life imprisonment or not. In the first case,…, and so on). Other representations are also possible, for example in the form of histograms as shown in Figure 2, for a hypothetical medical diagnosis decision system.

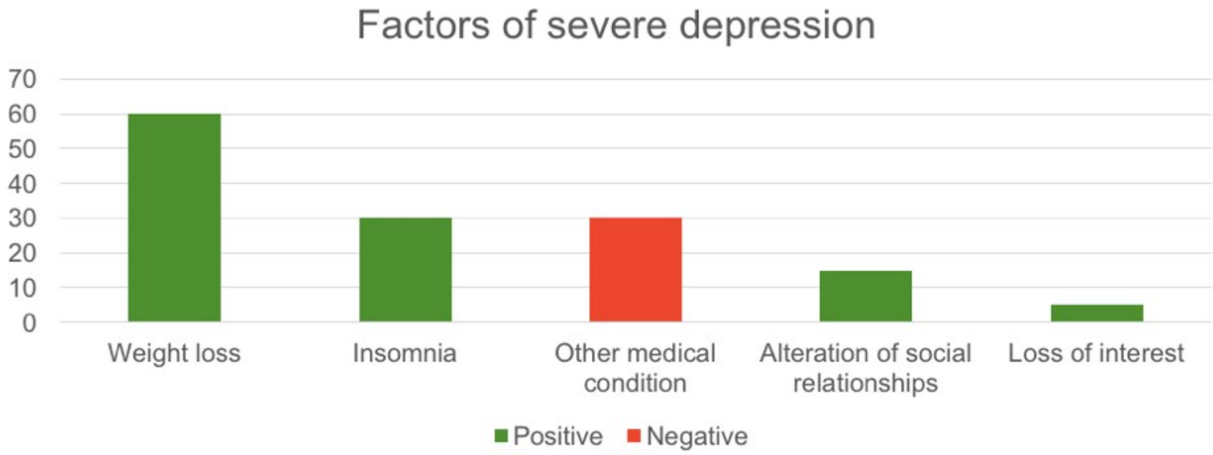
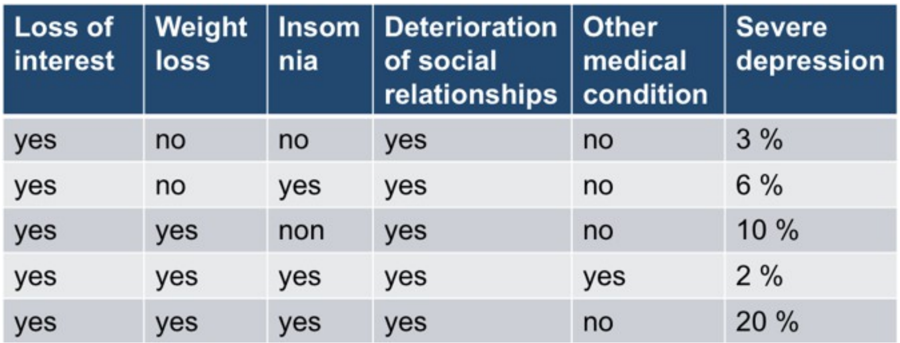


Figure 2

This type of histogram highlights the relative weights of the factors taken into account in a decision. For example, according to Figure 2, the main symptoms of depression are weight loss and insomnia. Deterioration in social relationships and loss of interest are also taken into account, but their respective weights are lower. The fact that the person concerned suffers from another medical condition is a negative factor in the depression diagnosis (because the symptoms may be caused by this other medical condition).

Another way of understanding the effects of combinations of factors on the results of an algorithm is to represent them as probabilities in explanation tables. For example, the table in Figure 3 shows that a depression diagnosis is suggested for 20% of patients presenting all the symptoms indicated as long as they do not suffer from another medical condition.



Loss of interest	Weight loss	Insom nia	Deterioration of social relationships	Other medical condition	Severe depression
yes	no	no	yes	no	3 %
yes	no	yes	yes	no	6 %
yes	yes	non	yes	no	10 %
yes	yes	yes	yes	yes	2 %
yes	yes	yes	yes	no	20 %

Figure 3

Other kinds of models, such as decision tables and different types of Bayesian networks, can be used to describe the logic of an algorithm. Since our aim in this paper is limited to providing an overall view of available options, we do not discuss them further. Interested readers can find more comprehensive presentations in the literature.[88]

4.2.2. Local explanations of specific results

Some people may find the global logic of an algorithm too complicated or too abstract. Another option, which is more specific and often more intuitive, is to provide explanations focused on particular cases. For example, the LIME system[89] makes it possible to identify the parts of an image that have mostly contributed to the identification of a pattern. As an illustration, its authors show[90] that a frog is identified in a picture with a probability of *0.54* by highlighting its head and as a billiard table with a probability of *0.07* by highlighting several small red balls (which in fact represent the tips of the figurine's front legs). Similarly, some methods provide as explanations the main factors (for example, symptoms or family history) justifying a decision for a given case.

Another approach to provide explanations consists in generating, for a particular case, other cases leading to the same decision, or close cases leading to different decisions. For example, a prospective student whose university application has been rejected by the Admission Post-Bac[91] algorithm would be able to understand why he was rejected if he could get profiles that were similar to his own but which have been accepted. This type of explanations is sometimes called counterfactual.[92]

4.2.3. General remarks: complementarity and variety

The different modes of explanation described above have their respective advantages and limitations. For example, local explanations can address the needs of a person who wants to understand the reasons for a decision affecting him, but they would not be sufficient for a supervisory authority wishing to verify the legality of an algorithmic decision system. No explanation mode outperforms the others, because they can target very different people with varying technical background and motivations. These approaches are in fact complementary rather than competing and they could be used in conjunction. It is important also to emphasise that an explanation generally provides a simplified vision of an algorithm. Generally speaking, there can be a tension between the precision of an explanation and its intelligibility. Even an intuitive mode of representation like decision trees can lead to explanations that are difficult to understand if there are too many nodes or the order of criteria is not well-chosen. It may sometimes be necessary to simplify a representation (for example, by reducing the number of considered criteria) at the expense of a loss of precision. Ideally, an explanation system should provide interaction facilities. Some systems, such as Elvira[93], have been developed with this goal in mind. Elvira has a graphical interface allowing users to enter assumptions (for example, a symptom such as 'headache') and to observe the effects of this new information on the possible diagnoses. The relations between the entities (for example, symptoms and diagnosis) are represented in the form of a graph[94] and the user can, at any point, choose to show varying levels of detail into each node. We believe that this ability to interact with an explanation system, particularly by testing different assumptions, is a fruitful research direction for improving the intelligibility of algorithms.

4.2.4. Implementation of explanation systems

The AdFisher and Sunlight systems mentioned in Section 3.3.2 provide global explanations in the 'black box' mode, that is, without any knowledge of the text of the algorithm or code of the software implementing it. The LIME system discussed in Section 4.2.2 provides local explanations.

It is based on the idea of generating examples of inputs close to the value of interest in order to analyse the behaviour of the algorithmic system in its neighbourhood. This behaviour, which is generally simpler than the global logic of the system, can be translated into histograms (such as Figure 2) or explanations of images (as illustrated in 4.2.2) for example. When the code of the algorithm is available, another option is to try to analyse it to derive explanations. This analysis can be more or less difficult depending on the nature of the algorithm. When the algorithm is not based on machine learning techniques (like the Admission Post-Bac algorithm), it may be possible to apply program code analysis methods (which are well known in the area of programming language compilation), for example to extract dependencies between inputs and outputs. The situation is very different, however, for machine learning systems. The first reason, which is obvious, is that their results do not depend only on the code of the algorithm but also on the training data. In addition, while some techniques (such as Bayesian networks) are amenable to different types of analysis[95], others (such as neural networks) are more challenging. Research has nevertheless been carried out to analyse this type of algorithm, in particular to derive explanations in the form of rules or decision trees. However, they have several limitations. On the one hand, they are very dependant on the types of algorithms considered, which means that they have to be adapted if the algorithm is modified. On the other hand, they are generally not suitable for the explanation of deep neural networks.[96] For these algorithms, the knowledge of the code is not very helpful, and black box techniques remain the only possible option.

Rather than trying to explain the results of a system *a posteriori*, the ideal scenario would be to take into account explanability requirements during the development phase, thereby providing ‘intelligibility by design’. This approach can be useful both to allow developers to improve their algorithms and to facilitate the adoption of machine learning techniques in certain application areas in which opacity cannot be tolerated. A promising step in this direction is to make the learning algorithm produce not just its nominal results but also explanations of these results. For example, the ‘rationalisation of predictions’ system recently proposed by the MIT[97] allows users to generate the excerpts of a text that have mostly contributed to the results (prediction). These excerpts must meet two criteria: they must be interpretable (short and contiguous) and they must be sufficient to explain the result. In other words, an analysis of the text reduced to these excerpts must lead to the same result as an analysis of the whole text. This approach seems very promising and will probably lead to a new generation of explanation systems in the future.

5. Conclusion

This paper argues that focusing on transparency in the sense of making the code of algorithms publicly available is not the only option, and even not the best solution to make algorithmic decision systems more acceptable. A new approach of ‘accountability’, encompassing certification with respect to well-defined criteria and intelligibility seems to provide more interesting perspectives. However, it also raises complex questions, both legally and technically, especially in domains like health and justice. The explanations of an algorithmic decision system can take different forms depending on the domain concerned, the algorithmic techniques used (determinist or probabilistic, learning-based or not, supervised or not, etc.) and the target audience (for example, professional or individual). For algorithms relying on machine learning, explanations should cover not only the logic of the model but also the training data set. The objective is not to make the system ‘transparent’ in the sense that anyone would be able to see and understand the entire process but to provide sufficient information to allow its users and decision-subjects to challenge its results. The right to challenge a decision is especially important for court judgments. However, the quality of algorithmic decision systems in the justice sector depends on the availability of previous decisions (case law), which is not sufficient in many countries. In France, the *Loi pour une République numérique* (Law for a Digital Republic), adopted on 7 October 2016, should improve the situation[98], but its implementation and the difficulties related to anonymisation requirements could considerably delay the process.

This paper also points out some blind spots and gaps in the European and French legal framework regarding fairness and intelligibility of algorithmic decision systems. The personal data protection regulation, in Europe and in France, has a limited scope. It does not explicitly cover algorithmic decision *support* systems and it is dedicated to data subject protection. There is a need for a broader reflection. Most algorithmic software currently available in the medical and legal sectors are proposed to help, as decision support tools. How could such decision systems be truly useful if they may be biased and if the user does not receive information about the logic and the training data set used? In France and Europe, existing legal requirements are unsatisfactory regarding verification and explanation. Professional users should be entitled with a right to receive more information. In the medical domain, it is necessary to strengthen the legal framework on the certification of medical devices. In the justice sector, general principles may not be sufficient to face the challenge raised by algorithmic systems. In our view, there is an urgent need for a new specific regulation. This is necessary to protect not only data subjects, but more broadly patients, litigants and respondents. They would directly benefit from a right to know that an algorithmic decision system has been used, and they would undirectly benefit from a better informed medical or judicial decision. Of course, these suggestions have to be discussed and a public debate is imperative, as the issues at stake are complex and far-reaching. Beyond the domain of health and justice, it could be difficult to identify categories of situations in which verification and explanation should be required for the use of algorithmic decision systems. The challenges are also varied and numerous on the technical side. As discussed above, different types of explanations can be provided to different types of stakeholders and these explanations should meet two potentially conflicting goals: intelligibility and precision. Measuring and ensuring these two goals is a major challenge. Explaining algorithms is a fast-growing research area and much progress is to be expected in the next decade. To be successful, this research should involve a variety of technical expertises and backgrounds including in particular artificial intelligence, software, and human-machine interaction.

[1] CNRS, Université de Nantes, DCS UMR 6297, Faculté de Droit de Nantes Chemin de la censive du tertre, BP 8130744 313 Nantes Cedex 3, France.

[2] Univ Lyon, Inria, INSA Lyon, CITI, F-69621 Villeurbanne, France.

[3] We use the expression `algorithmic decision systems' to denote both automated decision-making and decision support systems. This distinction is further discussed in Section 2.

[4] Ananny M. and Crawford K. (2016), 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', *New Media & Society* Vol. 20, 3: 973.

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Article 5(1).

[6] Loi n° 78-17 du 6 Janvier 1978 (on Information Technology, Data Files and Civil Liberties), amended: Data Protection Act 1978 (last amended in June 2018), Article 6(1): '*Processing may be performed only on personal data*

that meet the following conditions: 1. the data shall be obtained and processed fairly and lawfully'.

[7] See, for example, the report published by the French General Council for Economy, Industry, Energy and Technology: Modalités de régulation des algorithmes de traitement de contenus, 13 May 2016.

[8] Diakopoulos N. (2016), 'Accountability in algorithmic decision making', *Communications of the ACM*, Vol. 59, No2: 56. Binns R. (2018), Àlgorithmic accountability and public reason', *Philosophy & Technology*, Vol. 31, Issue 4, pp 543-556.

[9] Whitehead N.(2014), 'IBM Supercomputer Tackles Brain Cancer', *Science Magazine*, 20 March: http://www.sciencemag.org/news/2014/03/ibm-supercomputer-tackles-brain-cancer (last accessed:May 2018)

[10] Ricoeur P. (2011), *Le Juste*. 2 (Paris : Editions Esprit).

[11] Torasso P. (1985), 'Knowledge based expert systems for medical diagnosis', *Statistics in Medicine,* Vol. 4 : 317; Holman J.G. and Cookson M.J. (1987), 'Expert systems for medical applications', *Journal of Medical Engineering & Technology*, Vol. 11, Issue 4 :151 ; Korpinen F.H (1993), 'Sleep Expert--an intelligent medical decision support system for sleep disorders', *Med Inform*, Apr-Jun;18(2) : 163 ; Cléret M., Le Beux P. and Le Duff F. (2001), 'Les systèmes d'aide à la décision médicale', *Les Cahiers du numérique*, Vol. 2, N° 2 : 125.

[12] Thomasset C. and Vanderlinden J. (1998), 'Cantate à deux voix sur le thème "Une révolution informatique en droit" ?', *Revue Trimestrielle de Droit Civil* 1998 : 315; Bourcier D. (2007), 'À propos des fondements épistémologiques d'une science du droit', in Aguila Y. (ed.), *Quelles perspectives pour la recherche juridique?* (Paris : Presses Universitaires de France).

[13] French Public Health Code : Code de la Santé Publique, Article L. 4131-1 et seq. Please note all quotations from French sources without published English translations have been translated into English for the purpose of this article.

[14] French Public Health Code : Code de la Santé Publique, Article L. 4161-1.

[15] See, for example, in French jurisprudence: Cour de Cassation, Criminal Chamber, 11 January 2012, Appeal No 10-88.908, which acknowledges the fact that an optician using a non-contact air-puff tonomotor to measure intraocular pressure may be prosecuted and convicted for an illegal practice of medicine.

[16] Medicine cannot, moreover, be practised under a pseudonym (Code de la Santé Publique, Article L. 4113-3).

[17] https://www.conseil-national.medecin.fr/sites/default/files/code_de_deontologie_version_anglaise.pdf (last accessed: May 2018)

[18] French Code of Civil Procedure: Code de Procédure Civile, Article 12.

[19] French Data Protection Law: Loi Informatique et Libertés, Article 9: '*personal data processing that relates to offences, convictions and security measures can only be implemented by: 1) jurisdictions, public authorities and legal persons managing a public service and acting within the framework of their legal powers; 2) judicial assistants, strictly for the purposes of carrying out the tasks entrusted to them by the law*' as well as by legal persons acting on behalf of the victims of certain legally listed infringements. Article 9 amendment by the June 20, 2018 Act (*Loi n°2018-493 du 20 juin 2018*), amending the Loi Informatique et Libertés, has been cancelled by decision of the French Constitutional Council (*Conseil constitutionnel n° 2018-765 DC du 12 juin 2018*).

[20] Directive 95/46/CE of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Article 15.

[21] GDPR, Article 22.

[22] Since Regulation 2017/745 of 5 April 2017 came into force, this also concerns 'making available on the market or putting into service' as well as 'clinical investigation(s)' (Article 1, 1).

[23] Council Directive 93/42/EEC of 14 June 1993 concerning MDs (OJL 169 of 12 July 1993, p. 1).

[24] Regulation 2017/745, Article 120.

[25] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (OJL 117, 5 May 2017, p. 1).

[26] See also, the interpretative guidance document: EU MEDDEV 2.1/6 guide provides 'Guidelines on the qualification and classification of stand alone software used in healthcare within the regulatory framework of medical devices' (January 2012).

[27] Directive 93/42, Article 9; Regulation 2017/745, Article 51. Active implantable medical devices (AIMD), formerly regulated by a specific legal instrument, have been integrated into Class III subject to their own specific regulations.

[28] http://ansm.sante.fr/var/ansm_site/storage/original/application/a97a735f392805d3e9a927dfe2514aba.pdf (last accessed: May 2018)

[29] http://ansm.sante.fr/Activites/Mise-sur-le-marche-des-dispositifs-medicaux-et-dispositifs-medicaux-de-diagnostic-in-vitro-DM-DMIA-DMDIV/Logiciels-et-applications-mobiles-en-sante/(offset)/1 (last accessed: May 2018)

[30] http://ansm.sante.fr/Activites/Mise-sur-le-marche-des-dispositifs-medicaux-et-dispositifs-medicaux-de-diagnostic-in-vitro-DM-DMIA-DMDIV/Logiciels-et-applications-mobiles-en-sante/(offset)/1 (last accessed: May 2018)

[31] European Court of Justice, Fourth Chamber, 7 December 2017, Case C-329/16, *Snitem & Philips v. France*.

[32] Under the aegis of Directive 93/42, a diagnostic aid software that involves no measuring or functioning directly linked to another MD falls into DM Class IIa. A diagnostic aid software coupled with a measuring device has the same IIa classification but with a specificity relating to the compulsory intervention of a notified body to certify the metrology. A diagnostic aid and intervention software involving direct therapeutic action must be categorised as Class III, which calls for systematic recourse to verification by a notified body.

[33] Directive 93/42/EEC, Article 1.

[34] Between 2002 and 2018 in France, 387 records of registrations with the ANSM mention 'software', but it is not possible to identify the declarations of Class IIa MDs between 2002 and 2010. Since 2010, when a list of the registrations including Class IIa MDs became available for consultation, there have been 358 registrations of 'software', including 276 in Class IIa, 76 in Class IIb and 10 in the AIMD Class (http://ansm.sante.fr/Activites/Mise-sur-le-marche-des-dispositifs-medicaux-et-dispositifs-medicaux-de-diagnostic-in-vitro-DM-DMIA-DMDIV/DM-classes-IIa-IIb-III-et-DMIA-Communication-et-liste/(offset)/4) (last accessed: May 2018).

[35] Regulation 2017/745, Annex VIII, Rule 11.

[36] Directive 93/42, Article 3 (essential requirements); Regulation 2017/745, Article 5 (general safety and performance requirements).

[37] Directive 93/42/EEC, Article 5.

[38] Directive 93/42, annex 1; Regulation 2017/745, Annex 1, chapter 1, Point 1.

[39] Regulation 2017/745, Annex 1, Point 14.4.

[40] Regulation 2017/745, Annex 1, Point 14.5

[41] Regulation 2017/745, Annex 1, Point 15.1 (designed 'in such a way as to provide sufficient accuracy, precision and stability […] based on appropriate scientific and technical methods').

[42] Regulation 2017/745, Annexe 1, Point 17.1 ('designed to ensure repeatability, reliability and performance' and, in the event of a fault, put in place 'appropriate means […] to eliminate or reduce as far as possible consequent risks or impairment of performance').

[43] Regulation 2017/745, Annex 1, Point 17.4.

[44] http://www.astm.org/Standards/forensic-science-standards.html (last accessed: May 2018)

[45] This has been the case in France, since the introduction of law no 94-653 and 94-654 of 29 July 1994 relating to respect for the human body, for DNA identification (Article 16-11 of France's Civil Code, Article 226-28 of France's Penal Code) and for the examination of genetic characteristics (Article L. 1131-2-1 of the Code de la Santé Publique).

[46] Renard B. (2007), 'Les analyses génétiques en matière pénale: l'innovation technique porteuse d'innovation pénale?', Champ pénal: http://champpenal.revues.org/1241; DOI: 10.4000/champpenal.1241 (last accessed: May 2018)

[47] Standard Guide for Establishing Confidence in Digital Forensic Results by Error Mitigation Analysis, Active Standard ASTM E3016.

[48] Sayn I. (ed.) (2002), *Un barème pour les pensions alimentaires?* (Paris : La Documentation française); Sayn I. (ed.) (2014), *Le droit mis en barèmes?* (Paris : Dalloz) ; Service des Affaires Européennes et Internationales du Ministère de la Justice (2013), Etude Juriscope sur les référentiels d'indemnisation des préjudices extrapatrimoniaux faisant suite à une atteinte corporelle en Allemagne, Angleterre et Pays de Galles, Belgique, Espagne, Italie, Pays-bas, Report, Paris :

June 2013.

[49] For a clear example of this case law: Civ.2, 22 November 2012, Appeal no 11-25988.

[50] Conseil d'Etat, 7 December 2015 Application nos 378323, 378325, 386980; Conseil d'Etat, 6 January 2016 Application nos 388860, 388771, 390036; Conseil d'Etat 27 July 2016 Application no 390119; Conseil d'Etat 17 October 2016, Application no 400375.

[51] See, for example, Cour Administrative d'Appel de Lyon, 20 February 2014, Application no 13LY00269.

[52] Cadiet L. (2018), *L'open data des décisions de justice - Mission d'étude et de préfiguration sur l'ouverture au public des décisions de justice*, Ministère de lajustice, Janvier 2018: http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/184000019.pdf (last accessed: May 2018)

[53] APREF, Report 'Indemnisation des dommages corporels: Analyse et perspectives', June 2013: '*The development of tools therefore seems to be the culmination of practices validated by the jurisdictions. On this matter, the insurers are undeniably a step ahead regarding the use of a unique medical compensation scoring table, which is imposed on the jurisdictions. The professional journal Concours Médical's indicative compensation scoring table, which is frequently updated, has been the blueprint for the development of the expert assessment missions*':

http://www.apref.org/sites/default/files/espacedocumentaire/apref_note_indemnisation_des_dommages_corporels_2013.pdf (last accessed: May 2018)

[54] Vergès E., Vial G. and Leclerc O. (2015), *Droit de la preuve* (Paris: Presses Universitaires de France).

[55] Folllowing a certain period of time on a court of appeal list, the expert can ask to be entered onto the national list.

[56] Decree no 2004-1463 of 23 December 2004 on legal experts.

[57] Article R. 221-9 Code de Justice Administrative.

[58] For civil law: Article 232 Code Civil. For criminal law: Article 156 Code Procédure Pénale. For administrative law: Article R. 621-1 Code de Justice Administrative.

[59] Meimon Nisembaum C. (2007), 'Barème ou référentiel: du pareil au même', *Reliance* 1(23) : 97 ; Schmitzberger-Hoffer V. (2017), 'Barèmes d'indemnisation: toujours plus lourde, la charge de la preuve…', *Gazette du Palais* 7 feb. 2017, No 6 : 40.

[60] See, for example, the ministerial response from the Justice Minister dated 20 June 2006 (J0 of 20/6/2006): '*In view of the improvement in victim compensation terms, the Chancellery endeavours in particular, among the studies currently underway, to implement the means for standardising case law without detracting from the judge's discretion. In light of this objective, the possibility of putting in place a system of reference of sums allocated by the courts of appeal in cases of physical injury is currently the subject of a detailed study*'.

[61] Douai and Rennes Courts of Appeal have tried a Predictice software.

[62] For France, see in particular law No 2008-496 of 27 May 2008, which brings a number of provisions into line with EU law on combatting discrimination. For the United States, see the Employment Act (1967), the Fair Housing Act (1968) and the Equal Credit Opportunity Act (1974).

[63] Starr S. B. (2014), 'Evidence-based sentencing and the scientific rationalization of discrimination', *Stanford Law Review*, Vol. 66. See also the EPIC website: https://epic.org/algorithmic-transparency/crim-justice/ (last accessed: May 2018).

[64] Pederschi D., Ruggieri S. and Turini F. (2012), 'A study of top-k measures for discrimination discovery', SAC'12/2012, Proceedings of the 27th Annual ACM Symposium on Applied Computing pp 126-131.

[65] That is, any algorithm that satisfies this criterion also satisfies the criterion of 'disparate impact'.

[66] Hajian S. and Domingo-Ferrer J. (2013), 'A methodology for direct and indirect discrimination prevention in data mining', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, N° 7.

[67] For example, by avoiding the use of factors that correlate with the prohibited factors.

[68] That is, they will be the least disrupted and therefore much more useful. It is even possible in this case to integrate minimisation of loss of usefulness as a constraint to be met by the system (see, for example: C. Dwork, M. Hardt, T. Pitassi, O. Reingold, Innovations, Fairness through awareness, in *Theoretical Computer Science* (ITCS'12),

ACM, 2012).

[69] This is a major issue most notably in the case of search engines.

[70] A. Datta, M. C. Tschantz, A. Datta, Automated Experiments on ad privacy settings, *Proceedings of Privacy Enhancing Technologies (PETS)*, 2015.

[71] Google Ad Settings.

[72] M. Lecuyer, R. Spahn, Y. Spiliopoulos, A. Chaintreau, R. Geambasu, D. Hsu, Sunlight: fine-grained targeting detection at scale with statistical confidence, *Proceedings of CCS'15*, 2015.

[73] GDPR, Article 14, 2., g).

[74] GDPR, Article 15, 1., h).

[75] Wachter S., Mittelstadt B. and Floridi L. (2017), 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation', *International Data Privacy Law*, Vol. 7, N° 2: 76. See also Goodman B. and Flaxman S. (2016)*,* European Union regulations on algorithmic decision-making and a 'right to explanation': https://arxiv.org/pdf/1606.08813.pdf (last accessed: May 2018)

[76] Malgieri G. and Comand G. (2017), 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation', *International Data Privacy Law*, Vol. 7, N°4: 243.

[77] Decree no 2017-330 of 14 March 2017 on the rights of persons who are the subject of individual decisions made on the basis of algorithmic processing.

[78] CERNA (2017), Ethique de la recherche en apprentissage machine, June 2017.

[79] CNIL (2017), Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle, Synthèse du débat public animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, décembre 2017.

[80] Resolution of the European Parliament of 16 February 2017 containing recommendations to the Commission concerning the rules of civil law on robotics (2015/2103(INL)).

[81] European Court of Human Rights, 18 March 1997, No 21497/93, Mantovanelli c/ France: Rec. CEDH 1997, p. 424.

[82] For France: Code de Procédure Civile, Article 455.

[83] Conseil Constitutionnel, Decision no 2005-514 DC, 28 Apr. 2005.

[84] Cour de Cassation, 2010 Report:

https://www.courdecassation.fr/publications_26/rapport_annuel_36/rapport_2010_3866/etude_droit_3872/e_droit_3873/obligation_se_justifier_expliquer_3875/obligation_motivation_19404.html ; Louvel B. (2015), Working group report: https://www.courdecassation.fr/cour_cassation_1/reforme_cour_7109/cour_cassation_32575.html (last accessed: May 2018)

[85] Article 39(1) Loi Informatique et Libertés (French data protection act 1978 amended) provides, most notably, for the communication by the person responsible for processing personal data of 'information allowing persons to become familiar with and to contest the logic underpinning the automated processing in the cases of decisions taken on the basis of said processing and producing legal effects with regard to the data subject'. This is subject to restrictions, however, which are most notably aimed at protecting the intellectual property of the person responsible for the processing.

[86] Article 13(2) RGDP specifies that 'the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: […] the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject'.

[87] Wachter S., Mittlestadt B. and Floridi L. (2017), 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation', *International Data Privacy Law,* Vol. 7, No. 2.

[88] Freitas A. A. (2014), 'Comprehensible classification models – a position paper', *ACM SIGKDD Explorations Newsletter* Vol. 15 Issue 1 : 1.

[89] Ribeiro M. T., Singh S. and Guestrin C. (2016), '"Why should I trust you?" Explaining the predictions of any

classifier', *Proceedings of KDD 2016*.

[90] https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime (last accessed: May 2018)

[91] France's portal for processing applications for undergraduate university courses.

[92] Wachter S. and Mittlestadt and Russel C. (2017), "Counterfactual explanations without opening the black box: automated decisions and the GDPR", *Harvard Journal of Law and Technology*, forthcoming. Available at SSRN: https://ssrn.com/abstract=3063289 (last accessed: May 2018).

[93] Lacave C., Atienza R. and Diez F.J. (2000), 'Graphical explanation of Bayesian networks', *Proceedings of the International Symposium on Medical Data Analysis (ISMDA-2000)*. Springer-Verlag, Heidelberg, 2000, pp. 122–129.

[94] Which is in fact a Bayesian network.

[95] Lacave C. and Diez F. J. (2002), 'A review of explanation methods for Bayesian networks', *The Knowledge Engineering Review, 17*(2), pp. 107-127.

[96] Hailesilassie T. (2016), 'Rule extraction algorithm for deep neural networks: a review'*, International Journal of Computer Science and Information Security*, Vol. 14, No. 7.

[97] Lei T., Barzilay R. and Jaakkola T. (2016), 'Rationalizing neural predictions': https://arxiv.org/abs/1606.04155 (last accessed: May 2018)

[98] With the exception of criminal matters however.