The black box and the physician

Investigating the legal issues of explainability in medical AI

Sonia Desmoulin, Translated by **John Crisp**

Dans **Réseaux 2024/6 No 248** , pages 227 à 261
Éditions **La Découverte**

ISSN 0751-7971
ISBN 9782348085987
DOI 10.3917/res.248.0227
Date de mise en ligne : 14/01/2025

THE BLACK BOX AND THE PHYSICIAN

Investigating the legal issues of explainability in medical AI

Sonia DESMOULIN

Translated by John Crisp, LINC Languages OÜ

DOI: 10.3917/res.248.0227

THE BLACK BOX AND THE PHYSICIAN

Investigating the legal issues of explainability in medical AI

ABSTRACT

The deployment of decision-support systems based on artificial intelligence technologies has highlighted issues of explainability and transparency when automated processing renders the path from input data to output results opaque. By studying systems developed for the treatment of cancer and multiple sclerosis, this paper compares the theoretical thinking of legal experts with the practical concerns of medical professionals on this issue. Drawing on the legal literature and on interviews and evaluation forms, the analysis reveals a partial disconnect: professionals are far more concerned with training data and validation methods than with the internal "logic" of the computer algorithm. They seem to consider explainability as a factor for developing complementary (more or less transparent) tools, rather than as a choice to be made in order to eliminate one path. However, the adoption of the new European Regulation on AI does offer interesting prospects for convergence that is likely to give substance to the principles of explainability and transparency, and their legal variations.

Keywords: AI; explainability; transparency; law; oncology; neurology.

LA BOÎTE NOIRE ET LE MÉDECIN

Enquête sur l'enjeu juridique de l'explicabilité des IA médicales

RÉSUMÉ

Le déploiement d'outils d'aide à la décision basés sur des techniques d'intelligence artificielle a mis en lumière les enjeux d'explicabilité et de transparence dès lors que le traitement automatisé rend plus opaque le cheminement allant des données d'entrées aux résultats en sortie. À partir d'outils développés pour la prise en charge du cancer et de la sclérose en plaques, cet article propose de confronter la réflexion théorique des juristes et les préoccupations pratiques des professionnels de la médecine sur cette question. En s'appuyant sur la littérature juridique, d'un côté, et sur des entretiens et des fiches d'évaluation, de l'autre, l'analyse aboutit au constat d'une disjonction partielle : les professionnels se soucient bien davantage des données d'entraînement et des modalités de validation que de la « logique » interne de l'algorithme informatique et considèrent les options techniques plus ou moins explicables de manière complémentaire plutôt qu'alternative. L'adoption du nouveau Règlement européen sur l'IA offre toutefois des perspectives de convergence intéressantes susceptibles de soutenir l'effectivité des principes d'explicabilité et de transparence sous leurs diverses déclinaisons juridiques.

Mots-clés : IA ; explicabilité ; transparence ; droit ; cancérologie ; neurologie.

Tools that can automatically process digital data hold great promise in medicine, particularly when diagnosis is based on image analysis or biological tests[1]. This is the case for cancer and multiple sclerosis, for example, where the diagnostic and therapeutic decisions made by oncologists or neurologists are based on X-rays and/or biochemical, anatomical-pathological or microbiological analyses. The computational performance of artificial intelligence (AI) systems, now well known for their automatic pattern recognition and correlation spotting capacities, is expected to save time and enhance efficiency (Biancalana, 2023; Cerasa and Crowe, 2024). The prospect of such tools being incorporated into routine clinical practice in the near future seems all the more credible given that manufacturers in Europe and the United States have already obtained authorisation and certification to market medical devices with inbuilt AI in radiology and neurology (Muehlematter et al., 2021). However, although still sparse, studies on the integration of AI into medical practices show that its progress is not entirely smooth (Anichini and Geffroy, 2021) and that AIs are still a long way from becoming a regular presence in healthcare environments (Mignot and Schultz, 2022; Gillner, 2024). Most of these studies draw on experiments (Gaglio and Loute, 2023; Anichini, 2023) which, using different methods according to the context, demonstrate the wide variety of tools to which the "AI" label is applied and reveal the concerns of the players involved (Lombi and Rossero, 2024).

More often than not, the ethical and legal literature is therefore an exercise in anticipation, which leads to the formulation of a number of concerns, mainly relating to the degree of autonomy acquired by the machine. These include: the possibility and consequences of replacing doctors or certain specialties, particularly those most close engaged with the interpretation of measurements (Eon-Jaguin, 2022); the "dehumanisation" of the physician-patient relationship as a result of excessive reliance on "machine analyses"; issues of accountability, born either of the fear that healthcare professionals will be held liable for machine errors or, conversely, the possibility that doctors and healthcare institutions will cease to be accountable (Mazeau, 2018; Véron, 2023).

1. This article is the outcome of work carried out as part of two research projects: MALO (Machine Learning in Oncology, X. Guchet [dir.], INCA funding, no. 2021-146) and PRIMUS (Projection In Multiple Sclerosis, G. Edan [coord.], ANR-21-RHUS-0014).

Although there is as yet no global consensus on the definition of AI, an attempt at institutional harmonisation has been made (European Commission High Level Expert Group on AI, OECD), a venture that seems to have finally succeeded with the adoption of a European regulation to govern the marketing, commissioning and use of any "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Regulation [EU] 2024/1689: AI Act). This definition is intended for the purposes of this regulatory text and does not constitute a universal set of criteria. However, it does indicate the importance ascribed to the issue of autonomy. The interest shown in its adoption beyond the sphere of lawyers and industrialists could augur a gradual definitional alignment. So, while the term AI can continue to be used to refer to algorithmic tools that employ a variety of computing technologies (from expert systems to neural networks), the processes employed to discover correlations in a semi-supervised or unsupervised way lie at the heart of a new set of social, ethical and legal issues. The idea is gaining ground that "predictive machines do produce conclusions which, in a thousand and one ways, appear to be plausibly linked to the 'inputs', although the way in which they establish this link can nevertheless neither be anticipated, controlled nor monitored" (Benbouzid and Cardon, 2018, p. 19).

The opacity and transparency of these "black boxes" (Pasquale, 2016) have become key issues on the political and regulatory agenda, with widespread calls for transparency and explainability (Mittelstadt et al., 2016; Powers and Ganascia, 2020). In response to these demands, research efforts are being made in computer science to propose alternative approaches that are "explainable" throughout the stages from design to operation (Zhang et al., 2022).

For medical applications, the use of AI raises fears concerning a loss of critical thinking in the handling of results and concerning user disempowerment, as well as about the possible infringement of individual rights (Grote and Berens, 2020). Public authorities have made this concern a priority (WHO, 2021). In France, after asserting the need to maintain

oversight and intelligibility regarding individual administrative decisions based on machine processing, including in the healthcare domain (Article L. 311-3-1 of the Code des relations entre le public et l'administration – the law governing relations between the public and the administration), lawmakers introduced an obligation on healthcare professionals to provide information for the benefit of patients, as well as an obligation of "explainability" placed on designers for the benefit of users with regard to the "operation" of "algorithmic data processing trained on big data" (Article L. 4001-3 of the Public Health Code, arising from the Bioethics Act of 2 July 2021). At European level, the medical domain is directly targeted in the term "high-risk" applied to AI, since this category covers medical devices (MDs, except for class 1 MDs, which are not subject to a "third-party compliance assessment"), which entails an obligation of transparency and information sharing and a "right to explanation" (AI Act). These legal innovations form part of a legal and regulatory complex that also includes a medical device regulation (Regulation [EU] 2017/745, which refers to the "transparency" of the tools, but is relatively silent on information specifically about data processing) and rules relating to the protection of personal data. In France, following on from the provisions of the "Informatique et libertés" (data protection) act, the European regulation on the protection of personal data (GDPR: Regulation [EU] 2016/679) states that machine decision-making which uses personal data may only be legal under exceptional circumstances and is subject to certain requirements for the transmission of information on the "underlying logic" of the system.

These different sources of law give rise to legal questions about the scope of the transparency and explainability obligations, particularly as to what information and explanations must be provided, to whom, and what the consequences would be if it were technically impossible to supply explanations about the "underlying logic" of the system.

However, it may be asked whether the explainability of AI as discussed in the legal literature corresponds to the needs and concerns expressed by medical professionals. Regarding the ethical principle of explainability, a study dealing with the uses of AI in radiology noted: "The radiologists we met are far removed from these debates. In reasoning about purposes, they follow an experimental rationale. […] The radiologists we met evaluate the

software according to its results and correct what it suggests by reference to their own knowledge and experience" (Gaglio and Loute 2023, p. 162).

This suggests a disparity between, on the one hand, expectations for a proven result and, on the other, demands for transparency or explainable tools (London, 2019). There are three aspects to this possible discrepancy: it could indicate that the current legal debate, which is mainly led by digital market and data protection specialists, is ill-suited to the nature of medical activity, which entails a cost-benefit balance and specific evaluation systems; alternatively, it may reflect a blindness among certain professionals to new risks, with the onus being placed on the law to compensate for this lack of caution by imposing requirements on the designers, producers and users of AI; or thirdly, it could point to a difficulty in actually implementing the law if the stakeholders in the field do not understand the value of normative formulasthat are too far removed from their points of view. This third possibility may be seen in the context of the current dominance of professional regulation over a "state framework that is remote from actual practice" (Mignot and Schultz, 2022, p. 69). In an attempt to shed light, at least in part, on this question and these directions of thought, this article proposes to compare, on the one hand, the salient points of the discussions on the legal requirements for explanation and transparency applicable in France (French and European law) and, on the other hand, the elements that emerge from the opinions of healthcare professionals and medical AI developers in French cancerology and neurology projects.

For this study, the discussion of explainability in the legal field is informed by an analysis of texts and documents (published in French and English between April 2016 and June 2024), covering debates that emerged with the adoption of the General Data Protection Regulation (GDPR). However, it was not possible to gain access to the issues raised by medical and medical AI professionals through document analysis alone. While a few empirical studies in the humanities and social sciences have managed to capture the process of "building algorithms" (Jaton, 2020) or putting clinical practice to the test (Anichini and Geffroy, 2021), very few publications have focused on representations of the transparency or opacity of medical AI systems (Winter and Carusi, 2023), and even fewer in the French context. We therefore carried out a first analysis of a composite empirical body of material drawn from two types of sources – interviews and "questionnaire

forms". This material was collected within the framework of two research projects: first, the MALO (Machine Learning in Oncology) project, conducted between December 2021 and November 2024 under the direction of François-Xavier Guchet (philosopher) and funded by the Institut national du cancer (InCa), which aimed to identify the needs and values of cancer professionals and to examine the ethical and legal issues raised by the use of machine learning systems; secondly, the PRIMUS (Projection in Multiple Sclerosis) project, coordinated between January 2022 and April 2027 by Gilles Edan (neurologist) and funded under the ANR's (French national research agency) "University-Hospital Health Research" (RHU) initiative, which aims to design a tool to improve the management of multiple sclerosis patients by making it possible to reconstruct the trajectories of similar patients who have followed different treatments and to view the changes between the MRI scans of individual patients. Unlike surveys that look in depth at the specificities of a given field, this study focuses on cross-disciplinary issues and attempts to find common lines of thought across a variety of disciplines and projects. The aim is to compare general and theoretical legal perspectives with practical viewpoints, which means accepting the loss of some of the rich content of the context, which is always specific, in an attempt to make space for comparison. The interviews (n = 15) were conducted between March 2022 and March 2024, and were nondirective in format, lasting between one and three hours on average and undertaken by a multidisciplinary team containing members from the fields of epistemology, ethics, law and design[2]. The interview subjects were professionals from various specialties (cancerology, radiology, medical pathology, biochemistry, medical physics, adapted physical activity, bioinformatics and medical IT) working in the field of cancer in Lille, Lyon, Paris and Toulouse, either in cancer centres or medical informatics laboratories within healthcare establishments. The interviews focused on the career paths of the people involved, the tools (that might be) used in practice (such as software for detecting anomalies on mammograms) and those developed or under development, some on the initiative of healthcare professionals, others arising from collaborations instigated by commercial firms. Some of the interviewees were involved in several projects, others in just one. In addition to the interview data, the material

2. The team members were: Emanuele Clarizio, Sonia Desmoulin, Océane Fiant, Gaël Guilloux, François-Xavier Guchet and Alain Loute.

includes evaluation forms (n = 7) completed by neurologists charged with evaluating the prototype multiple sclerosis decision-support system: after a presentation and familiarisation session lasting a few hours, the neurologists manually answered 52 questions, including 40 closed questions (requiring a "strongly disagree" or "strongly agree" style response with five gradations) and 12 open questions. The evaluators also manually annotated some of the closed questions. These questions cover a range of topics, from operational issues (e.g. the user-friendliness of the interface or the relevance of the position in which a result is displayed) to operational factors (e.g. whether or not a result is displayed), or questions on validity or reliability (criteria for "trusting the results of a contextualisation") and on the information conveyed (missing or unnecessary information, the wish to be able to display the data source, etc.).

These interviews and questionnaires are limited in number and the interviewees are all involved in medical AI development projects. The results do not reflect the views of doctors whose practice has no connection with research. The contribution here is qualitative and concerns the expectations of medical professionals interested in the deployment of such tools. The answers collected did not focus on the problem of the explainability or transparency of medical AI tools, but the choice of non-directive interviews as the research method made it possible to confirm that these issues did indeed emerge without any specific prompting (Duchesne, 2000). The same applies to the interpretation of the questionnaires for the evaluation of the prototype in the PRIMUS project. The analysis of this varied body of material therefore had a dual purpose: to verify the presence (or absence) of concern about the high degree of opacity of certain computer technologies; and to identify different forms of concern expressed about the issues of understanding and information in these varied sources. I was starting with a question that was initially legal – should non-explainable AI be excluded? Should new information requirements be introduced? – but the interviews and evaluation sheets led me to discover what appeared to be strategies for designing and testing innovations intended for routine clinical use. The legal questions could then revert to the correlation between these expectations and strategies, on the one hand, and the factors highlighted by the law, on the other. The aim was thus to explore the meaning and scope of the new requirement of explainability in terms of its capacity to provide useful guidance for practice.

After presenting the salient features of the legal debates on explainability (1), the viewpoints of the healthcare professionals and medical AI specialists interviewed will be presented (2), in order to identify the issues around any divergences in the concerns expressed (3).

EXPLAINABILITY AS A SUBJECT OF LEGAL DEBATE

Do data subjects have a right to an explanation about machine decisions that affect them?

The GDPR was the first text to have prompted real discussion about whether or not explanations should be provided on the type of processes employed in arriving at an automated decision (Rochfeld, 2018). It has therefore given rise to the most numerous analyses. Indeed, in the GDPR, a general principle of transparency (Article 5, a]) constitutes the basis for requirements concerning the information that is to be given to people whose data is processed (Netter, 2020). A major debate has developed around the existence of a right to explanation with respect to algorithmic decision-making. It concerns the interpretation of Articles 13, 14 and 15, which respectively provide for information to be given to the person whose data is collected and the data subject's right of access to their processed data. Article 13(2) states that the subject of the data must be informed of "the existence of automated decision-making […] 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". Article 15 goes on to state that the data subject "shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed", and, where they are, to be informed of "the existence of automated decision-making […] 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". In the specific context of fully automated decisions (or decisions deemed to be so because the human component is negligible), the majority interpretation is that "by providing for information on the 'underlying logic' of the processing, the legislator is in fact requiring information relating to the 'reasoning', or in other words to the algorithm" (Huttner, 2022, p. 516). On this view, the right to information thus relates specifically to the

technical choices implemented, and this information is part of an explanatory process. Indeed, Interpretative Recital No. 71 of the GDPR provides that "processing [which is automated and the exclusive basis of a decision] should be subject to suitable safeguards, which should include specific information to the data subject and the right […] to obtain an explanation of the decision reached after such assessment and to challenge the decision"[3]. Consequently, the right to information here seems to be coupled with a right to explanation (Goodman and Flaxman, 2017; Malgieri and Comand, 2017). However, this interpretation is contested. Several arguments have been put forward to suggest that there is no such thing as a "genuine right to explanation" (Wachter et al., 2017). The authors point out that under Article 13 information is provided prior to the decision and cannot therefore relate specifically to the outcome. Conversely, Recital 71 raises the possibility of envisaging retrospective information, but is not binding. This is compounded by the fact that Articles 13 to 15 only provide for general information on the underlying rationale, without requiring any particular explanation of the specific decision complained about. From this perspective, explainability is discussed in terms of a "right to explanation", itself characterised in relation to a more general "right to information" with respect to the means available to the person targeted by the automated decision to enable him or her to understand and challenge it (Edwards and Veale, 2017; Wachter et al., 2018; Casey et al., 2019).

Can healthcare professionals claim a right to explainability?

In the medical field, the purpose of AI is to prepare and support decisions, since full automation cannot be envisaged without infringing medical ethics (Véron, 2023). Consequently, the above-mentioned articles are not an obvious resource, as they relate to decisions based exclusively on automated processing and they grant rights to the people whose data are processed (Bernelin and Desmoulin, 2021). On the other hand, some

3. Legal texts, particularly European directives and regulations, are introduced by preliminary comments called "Recitals", whose role is to help readers understand the meaning of the provisions, without having the binding normative value of the articles. Judges, in particular the Court of Justice of the EU, often refer to them to justify an interpretation of the text in the event of a dispute.

authors consider that Article L. 4001-3 of the French Public Health Code creates a genuine right to explainability for the benefit of the physicians who use machine processing (Huttner, 2022; Eon-Jaguin, 2022; Borrillo, 2023). However, several factors limit the scope of this opinion: the evasive wording of the text ("the designers of algorithmic processing mentioned in I shall ensure that its operation is explainable to users"), its restricted scope (algorithmic data processing where machine learning has taken place by means of big data) and the fact that implementation of the law is subject to an implementing decree. The development of a discussion in the legal literature thus appears to be dependent on this still pending supplementary text and on the proven existence of tools that correspond to the scope of application as defined.

The duty of transparency and the right to explanation in the AI Act?

The debate is likely to resurface with the adoption of the AI Act (Bensamoun, 2023), since the issue of explainability is asserted there as an element of the principle of "transparency" (Recitals 27 and 72). A comparison with the French text shows that there is scope for new discussion around AI medical devices (MD). The new regulation explicitly creates a "right to explanation of individual decision-making" for the benefit of the person affected by a decision based on a high-risk AI output and a corresponding obligation on the "deployer" (Article 86). On the other hand, under the "transparency" obligation, the Act sets out a twofold requirement. Firstly, high-risk AI must be designed and developed in such a way that its operation is "sufficiently transparent to enable deployers to interpret a system's output and use it appropriately" (Article 13.1). Secondly, information is provided by means of a user manual that sets out the AI system's characteristics, capabilities and performance limits, including the purpose of the system, its level of accuracy, robustness and cybersecurity, the foreseeable circumstances associated with its use that may give rise to risks, and also, "where appropriate", in particular "the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output", "specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk

AI system" and "information to enable deployers to interpret the output of the high-risk AI system and use it appropriately" (Articles 13.2 and 13.3, b]). Admittedly, a deployer is defined in the AI Act as "a natural or legal person, public authority, agency or other body using an AI system under its authority" (Article 3), which does not allow users acting under the authority of another person to be directly targeted. However, as doctors are ethically bound to independence, even when they are employees of a health establishment, it is theoretically possible to consider that this blind spot does not exist[4]. In the Artificial Intelligence Act, explainability is thus split into an obligation of transparency, in the form of specific information for the benefit of "deployers", on the one hand, and a right to explanation for the benefit of those who are the targets of decisions, on the other. However, the scope of the text is restricted to high-risk AI, which seems to cover class 2a, 2b and 3 AI medical devices in view of the conditions laid down[5], but could turn out to concern only some of them if the reference to the different levels of autonomy and to the possibility of post-deployment upgrades were ultimately interpreted in such a way as to exclude anything that is not related to unsupervised deep learning.

In the legal literature, the relationship between transparency and explainability lends itself to a variety of interpretations. It is possible to argue that they converge, since they are both part of the same concern with providing information to the various people concerned and with providing information that makes it possible to understand how a result (or a decision) is obtained from certain data. However, it is a challenge to clarify the specific nature of the connection between the requirement of explainability and the expectation of transparency, for several reasons: one might place the emphasis on the need to provide access to some understanding of the tool rather than insist on publication of the code (Desmoulin-Canselier and Le Métayer, 2018); or one might wish to show the need for a special obligation to provide information (G'sell, 2020; Eynard, 2020); or one might wonder

4. With regard to other personnel, article 4 is more vague on the obligations of suppliers and deployers.

5. In the regulation on MD, classes reflect the level of anticipated risk: class 1 represents a low risk (e.g. crutches), while class 3 is "potentially high-risk" (e.g. pacemaker). Medical software that qualifies as a medical device is, in principle, class 2a, i.e. "moderate or very moderate risk", other than in exceptional cases.

whether there is a right to explanation concerning the operation (or logic) of the tool and its beneficiary (Castets-Renard, 2018; 2020; Kaminski, 2019). The European and French texts offer much ground for debate, since they seem: sometimes to distinguish between the "right to explanation" of the people affected by the decision (GDPR and AI Act) and the right to information of the deployers (which does not exist in the GDPR but does in the AI Act); sometimes to affirm an obligation of explainability for the benefit of users that is distinct from a special obligation to provide information for the benefit of patients (French Public health Code); sometimes to construct "transparency" as a general principle (encompassing the provision of explanation) or a special obligation distinct from the right to explanation (AI Act). Quite apart from the subtleties of the distinctions, these approaches serve different purposes: for a person to gather information in order to be able to challenge a decision that has been taken against them; or for a person to have access to information in order to ensure that they are complying with their ethical obligations and making good use of the tool. In addition, the requirement of explanability – whether it takes the form of an enhanced obligation to provide information or an obligation to include mechanisms that facilitate challenges – raises questions about the legality of using a system based on technologies that are too opaque and incompatible with the possibility of explanation (Huttner, 2022). In requiring "sufficiently transparent operation", does the AI Act close the door on the use of decision support tools founded on deep learning? From this perspective, should we conclude that explanatory reverse engineering technologies are now indispensable (Wachter et al., 2018)? Should the scope of the term "sufficient" be assessed in the light of how a tool is used in the sphere of activity concerned and, in that case, could medical AI be considered as fulfilling regulatory requirements if it provides information on the design conditions, purposes, training data and validation conditions from a risk/benefit perspective specific to health products? Ultimately, however, these legal questions might not align with the concerns of the people who develop and want to use algorithmic tools.

EXPLAINABILITY AS AN EXPECTATION OF MEDICAL PROFESSIONALS

Aspects of the encounter between healthcare AI and research AI

Doctors sometimes come into contact with AI when clinical practice devices, such as a machine purchased by the radiology department, are enhanced with new data processing capacity, for example an automated image analysis option (Mignot and Schultz, 2022). The interviews do, however, provide evidence of their active role in the development of AI. The innovations mentioned are numerous: automatic detection of "normal" mammograms so that radiologists process only those images that show an abnormality; detection of tumour tissue (histopathological diagnosis support) from samples taken from the patient in the operating theatre; AI-based organ contouring to assess the patient's body composition from scanner images (to identify prognostic and predictive factors for toxicity risk based on muscle surface); extraction of data from anatomical pathology slides; automatic image annotation to train algorithms to recognise different types of cells and biological elements; structuring and sharing of medical file data to create cohorts of patients with characteristics or tumours of interest; assessment of clinical case complexity for referral to multidisciplinary "complex" case meetings or reminders of good practice recommendations for other types of case. The involvement of doctors varies: sometimes they are the initiators of the project (or major players in it); sometimes they are asked to assist with improvements to a tool that has already been developed. Doctors may only be asked to provide data (training data, test data) and expertise (annotation, validation). This expertise is provided both upstream – to determine expectations, criteria and relevant data during the design phase – and downstream, for example to evaluate a prototype. Depending on the context, therefore, the relationship with the instrument will be different. The discipline also plays a role, as some specialists, such as pathologists, focus on analysing tumours, while others, such as oncologists or physical activity consultants, interact with patients and treat pathologies as a human' complex. From one person to the next, and from one project to the next, the positions taken vary greatly: a radiologist takes part in discussions on design choices, while sticking to her area of expertise; a pathologist is totally involved in the co-construction of a tool for the automatic annotation of histology slides; a doctor has trained in

coding for the development of a tool that identifies cancerous cells, etc. As for the neurologists, they completed the questionnaire from the perspective of their clinical expertise. In these circumstances, it is understandable that the views expressed are not uniform. Nevertheless, it is possible to identify a number of value systems and representations that reveal attitudes to the issue of opacity and the strategies that need to be implemented to ensure that algorithmic tools find their place in medical practice.

The opacity of machine learning

The professionals interviewed share a picture of AI systems as statistical tools, which rely on a process of "simplification" ("simplified patient, simplified case, decision support" [medical informatics professor, 4 May 2022, Paris]) or on the quest for the "shortest path": "deep learning […] finds correlations and finds the shortest path possible" (pathologist, 25 October 2022, Toulouse). Medical AI is reintegrated into a familiar framework, one of medical practices that are already marked by technical innovation and "evidence-based medicine" (Morley et al., 2020), i.e. approaches based on measurement and numbers. It is a tool with a simplified vision of reality, like any instrument that seeks to capture a dimension of the organism or of a pathology. In the interviews, medical AI loses its specificity, whether it is situated as an extension of the patient record (a well-known tool, which already structures data and represents reality) or compared with medication (another tool that medical professionals have been working with for a long time). However, the technical leap represented by "machine learning", particularly when the results are not supervised, is not unacknowledged. It is here that concerns about information and understanding are expressed, echoing the questions raised in the ethical and legal literature.

This questioning of what goes on inside the algorithmic box is expressed by reference to the tools directly available to doctors. Referring to mammography interpretation software, one radiologist compares a machine that is "more open" with one that is not:

> "The algorithms are not the same. TX [software that helps to interpret digital mammograms in order to detect suspicious masses and determine whether they are malignant tumours], we don't really know. The algorithm

> remains very, very opaque, but there's also a problem of industrial property, so we can't go very far in understanding the algorithm either. T [a tool that interprets mammography scans, detects microcalcifications and suspicious lesions and assigns a risk score] was a little more open, because they immediately showed us that they were going to work with a probability of malignant lesion based on the images in their database. In other words, if [T] shows you, circles an image and gives you a figure – say 63%, this means that, in their database, in 63% of cases, that image was cancer" (radiologist, 2 May 2022, Lille).

Similar doubts are expressed about AI tools where a doctor has followed developments, but is only involved in supplying data. In the case of a device for analysing histology slides, one pathologist said:

> "We worked with the company [O]. They have a CHOWDER-type algorithm [convolutional neural network] which seems to be fairly official, but we don't know exactly… It's often black boxes, it's unsupervised: we don't really know what it's up to" (pathologist, 17 February 2023, Lyon).

Even in a case where the pathologist is more directly involved in the design of the automatic slide annotation tool, doubts arise:

> "There's sometimes a crazy side to deep learning when you say: this is brilliant, it's beyond anything. It beats everything in terms of performance, but how it works and how many decisions are ultimately based on small biases here and there that add up and tip the decision one way… It's hard to quantify" (pathologist, 7 September 2022, Toulouse).

Although the term transparency is not used frequently, or is used in a different sense, to which we will return, the non-directive interviews spontaneously conjure up images of "black boxes" and technology that is alternatively "more open" or "opaque". These expressions, which are warning signs, are not, however, associated with an attitude of pure and simple rejection, but relate more to questions about the conditions needed to build trust.

By contrast with some legal analyses, the fact that deep learning technologies are opaque does not lead to the idea they need to be rejected, but stimulates a quest for strategies to limit the blind spot and its effects. At least

five approaches can be identified in the interviews and questionnaires. As the qualitative analysis of the corpus did not reveal any obvious classification, they will be presented here in order of the stages of development, from the determination and design of a technology, through data selection, additional experiments and product testing, up to implementation projects. Depending on their specialty, their position in a project or even their affinities, professionals may refer to one or other (or several) of these stages. From the point of view of legal considerations, however, it is interesting to note that these options can be combined.

Limiting the effects of opacity

Algorithm design

It would be tempting to think that the physicians involved in AI development invariably engage in in-depth discussions with the data scientists over the choice of the technology, in particular on the basis of the existence of "open" or "explainable" technologies. However, the interviews do not provide an unequivocal answer on this point. Two medical professionals who collaborated on the development of an automated organ contouring tool for assessing a patient's body composition from scans convey opposing perspectives:

> "The data scientist who does machine learning or manifold learning, I don't go anywhere near that. […] That's not my field at all" (oncologist, 17 February 2023, Lyon).
> "We discuss the algorithms and the results, and so they have suggested several different algorithmic approaches for analysing the data" (physical activity researcher, 24 March 2023, Lyon).

These contrasting views may reflect personal affinities (for IT or for multidisciplinary teamwork) or work habits linked to the division of tasks (each specialty being considered best able to provide the best service). A medical AI system can also be made up of several combined tools which are applied differently. For example, the PRIMUS project combines two tools into a single system: the reconstruction of the trajectories of benchmark patients is a collaborative endeavour, with neurologists heavily involved in the technical choices; the analysis of brain images with the detection

of new tumours employs software already developed by a private partner company. This observation might put a different complexion on the pertinence of the frequent calls, particularly in the ethical and legal literature, for multidisciplinary approaches and IT training for doctors. The sense of working together is clearer when a data scientist moves into a pathologist's laboratory to develop an automated annotation tool for histological slides than when a pathologist is asked to annotate the slides in order to train a tool developed by a company producing innovative medical devices. The creation of a genuine "laboratory life" lies largely outside the legal framework.

Data production and verification

While not all the doctors questioned were directly interested in the choice of computer technology, they were all concerned about the data used. On this point, their concerns converge with the legal literature. Thus, the problem of technological opacity is partially addressed by verifying the quantity and quality of the data used to train and then test AI.

As a statistical tool, AI can only produce "robust" results if sufficient data are available. This can lead to adjustments, provided that the constraints of the tool can be reconciled with the medical approach. A question of this type was asked of the neurologists who completed the PRIMUS evaluation questionnaire: is it helpful for the future therapeutic scenario to be able to choose a treatment by class or molecule? One neurologist commented on the underlying issue: "We can deal with more patients if we just stick to the class" (neurologist evaluation questionnaire No. 1). In compiling the data by class or molecule, the designers chose not to present the results for the pharmaceutical speciality on the market, but to group the results of similar products together in order to gain in statistical robustness. This quantitative challenge is in fact so important that the respondents feel that if part of the database were inaccessible at the time of use, they would not use the tool.

The nature and source of the data used are also critical in determining the pertinence of the automatically generated projection. However, existing databases do not always meet the requirements. It therefore seems crucial

to produce the missing data, i.e. to collect good-quality data (in sufficient quantity), and to annotate or structure it, although this involves considerable effort. Whether the aim is to contour organs or to annotate histology slides: "It takes time" (medical physicist, 9 November 2022, Lyon); "it's thankless, it's complicated, it's time-consuming, and, what's more, there are quite often times when we're not really sure of what we're doing" (pathologist, 7 September 2022, Toulouse). This intense task of producing and verifying data that can be used by the machine but are also relevant to the medical framework of analysis plays a dual role in limiting the difficulties of understanding and evaluating the machine's performance. Indeed, it helps to elucidate the tool's medical logic and in creating the conditions for a positive benefit-risk balance. Using the right (well-chosen and well-prepared) data provides another level of insight into the tool.

Even with good-quality data, the problem of "small biases here and there" needs to be tackled in order to ensure that the system's opacity does not prevent errors in the results being spotted. Shared concerns were expressed, across all projects about the risk induced by biases in the data. With regard to the tool that automatically classifies breast cancer patient records to help identify complex cases, the example of the location of the lesion was mentioned:

> "What was on the right could be different from what was on the left. What was in one lesion could be different from what was in the lesion next to it. So each element had to be properly described in order to be relevant" (AP-HP project manager for medical IT, 24 February 2023, Paris).

Complementarity of algorithmic approaches (explainable/non-explainable)

In some projects, the opacity of the algorithm can also be partially overcome by combining unsupervised learning with more directly comprehensible approaches. For a project involving an automatic histology slide annotation instrument, the idea is that:

> "Ultimately, we get some interesting data by combining deep learning with more explainable methods" (pathologist, 7 September 2022, Toulouse).

In the case of the project for a tool to detect complexity in medical records, complementarity was sought using an older technology:

> "A machine learning algorithm was applied to a sub-sample, and then a decision rule was applied to that same sub-sample" (professor of medical informatics, 4 May 2022, Paris).

In the case of a project for a system for detecting the origin of tumours by analysing histological slides (search for the source organ in the event of the cancer spreading), the system is combined with an expert system:

> "So the first part is very machine learning, and the aim is to couple it to the expert system we were talking about earlier, where we have, basically, all the IHC [immunohistochemistry] literature" (cancer centre information systems director, 6 November 2022, Lyon).

These combinatorial approaches are the kind of experimental process that has been well described in the social sciences (Gaglio and Loute, 2023). However, it is notable that they are in conflict with a certain legal literature, which sees explainable AI as an alternative that needs to triumph over opaque AI.

Validation of medical AI

The devices used in medicine follow a variety of channels for verifying and demonstrating their effects: clinical trials, marketing authorisations (for medicines in the EU and US, and for medical devices in the US), certifications (MD certification, cyber-security certification, etc.), post-marketing tests (for inclusion in the list of products covered by social security, before purchase by a hospital department or for a post-marketing study). In particular, the producer must demonstrate the basis for the claims made, and the existence of a positive benefit-risk balance and a service rendered. These procedures thus correspond to the collective perceptions of a professional milieu where the paradigm is "evidence-based medicine". Regulatory authorisation or certification is not always enough to convince: scientific publications reporting the results and/or peer-to-peer trials complete the panoply of methods deployed to instil trust in the device. For the professionals interviewed, the choice of validation methods is a key element for

overcoming the doubts raised by the "black box" effect. The example of IBM's Watson for Oncology software, which was supposed to rival the best experts when it came to suggesting diagnoses and therapies based on large volumes of data, but which failed to convince, serves here as a deterrent:

> "Watson, we made them [the IBM team] come. We asked the World Scientific Director to come. I organised the meeting, and I remember it very well; I felt embarrassed for him. There were all the doctors working in molecular biology and so on, and I start by giving him four or five examples. Of the five, three were mistakes. In other words, the super-expert doctors knew that the suggestion was wrong" (oncologist, 17 February 2023, Lyon).

Watson for Oncology seemed to offer design guarantees – algorithmic know-how with theoretically proven results and collaboration with a team of expert clinicians (Memorial Sloan Kettering Cancer Center) – but failed to demonstrate its ability to provide accurate answers in real-life situations when tested in hospital departments, particularly outside the United States, notably because of its training biases (types of cancer, types of patients, forms of treatment proposed). Testing is therefore crucial, and must be done at every stage of development, from design to use, in a variety of conditions. The interviews have much to say about the different types of tests used to validate tools at different stages of their development:

> "They compared them with published neural networks, to get an idea of how they performed, etc. So they tested many different parameters, to see what was the best outcome between accuracy, error rate, analysis time, etc." (biochemist, 6 November 2022, Lyon).
> We're going to have to confirm that when a result is seen [with S], behind it, in anapathy, exactly the same thing is seen, and that false positive and false negative rates are calculated; in other words, that we calculate the specificity and sensitivity of the technology […] In sarcomas, it was really reassuring to be able to do retrospective clinical studies on samples that come out of the databases" (biologist, 29 March 2023, Lille).

For the decision support system in multiple sclerosis, the "main trust criteria" used in the evaluation forms, apart from the data sources and

quantities, are also "verification against data reported in publications" (neurologist evaluation questionnaire No. 3) and "validation of the software by the CRCs (expert centres)" (neurologist evaluation questionnaire No. 6). After the questionnaire-based prototype evaluation phase, a clinical trial is also planned. For these neurologists, as for the biologist taking part in the automated slide reading project in anatomical pathology, the benchmark remains the clinical trial, consonant with the habits ingrained since the advent of "evidence-based medicine" (Korica and Molloy, 2010). While in the eyes of legal experts, the issue of validation relates more to product safety and a positive benefit-risk balance, and therefore seems to need to be distinguished from questions about opacity, medical AI professionals clearly make the connection, considering that the issue of trust is central here and that this trust necessarily involves a confirmation procedure. This observation is similar to that made in the social science literature on the evaluation of medical device AI in radiology (Mignot and Schultz, 2022).

User information

Downstream of development, the transmission of information is the last factor cited by the professionals as needed to limit the effects of technological opacity. There is a certain convergence here with legal concerns, for example when it comes to informing users about training data. One oncologist considers that this issue is relevant to the different AI projects in which he is involved:

> "One of the other problems is that the physicians who are going to use these algorithms will need to know at least a minimum about the characteristics of the input datasets. Just as today they know the inclusion criteria for therapeutic trials, they will need to know what types of populations were included. […] My fear is that models will be used outside their training scope" (oncologist, 17 February 2023, Lyon).

The convergence is only partial, however, because when the notion of "transparency" is expressly mentioned, it refers to the idea that the doctor should find it easy to get to grips with the device or to learn about its precise performance. In the case of an organ contouring tool, for example, one researcher notes:

> "We tried to explain how we did things with our algorithm: what scores, such as the MAE in particular[6], we used to find out more or less how many percent or how far it is from the third lumbar vertebra. […] The link to the scientific article will be provided afterwards, but on every page of every patient record. […] We're really very much in favour of transparency" (physical activity researcher, 24 March 2023, Lyon).

Information serves both to help practitioners understand the AI device and to ensure its proper use within its intended "scope". As the recipient of the information is a professional, the argument implicitly seems to suggest that the information contains the seeds of explanation. This brings tacit knowledge into play (Anichini and Geffroy, 2021). The supposition is that knowing what data have been chosen and used for training helps people to understand what the AI tool is looking for and how it should be used. Moreover, the question of validation creeps into the argument: information is provided about the method of validation used even when the effort to explain is clearer, so the issue of opacity does not arise in its purely legal form. In the project for a multiple sclerosis decision support tool, an "Explanation of Results" section is included, and the questionnaire responses are all in favour of this option. However, this section is not designed to provide access to explanations of how the computer algorithm works, but rather to explain the projection through different ways of representing sources and results, as well as through information on the databases used.

CONVERGING PERSPECTIVES?

What lessons can be drawn from this comparison of legal and professional perspectives? The experts interviewed agree with the legal literature in warning against systems that use "machine learning" without the possibility of maintaining benchmarks to assess the processing performance. On the face of it, there is also a degree of terminological convergence, since the words information, transparency and explanation are employed. However, there are many disparities, both in terms of which issues are

6. MAE (mean absolute error) measures the mean differences between predicted and actual values.

deemed important and the meaning assigned to the wording. This divergence is partly due to the ambiguity of the concepts of explainability and transparency and the variety of ways in which they are used (Vuarin and Steyer, 2023).

The first notable difference concerns the idea, present in the legal literature, that certain technologies should be excluded simply because they are opaque. Although medical professionals express reservations about algorithmic tools that are considered to be highly opaque, they do not say that in consequence they should not be used. Instead, they deploy strategies to overcome the resulting blind spot: work on the relevant data, complementary technical approaches, validation tests and ways of providing information about the meaning of the result. This discrepancy might be attributable to the characteristics of professions whose history is punctuated with instrumental innovations and which accept uncertainty in the knowledge they teach. From a legal point of view, this raises the question of the relationship between health law and digital law. While the former has fully assimilated the specificity of medicine, by accepting risk in the benefit-risk assessment of health products and the construction of a largely fault-based system of medical liability[7], digital law sees medical AI as a set of systems that process sensitive data and that fall into a "high-risk" category. Medical law largely places the decision to take risks in the hands of the professional, guided by the relevant regulatory bodies (Comité consultatif national d'éthique [CCNE]; Haute Autorité de santé [HAS]), though the health authorities have only recently become involved in the evaluation of AI and still only to a limited degree (Mignot and Schultz, 2022). This raises the question of whether the issue of opacity arises in the same terms for medical AI systems as for others.

In some of the interviews, it is argued that "explainable" technologies can supplement or even support "opaque" technologies, but not replace them. The professionals interviewed consider that the explainable can be used to develop the opaque, and could even make the latter acceptable in different

7. A system of liability for defective products may also be used, but this presupposes that specific conditions have been met (enabling the professional to be considered a supplier) and is assessed differently depending on whether the tribunal is administrative (for hospitals) or judicial (for clinics or practices).

ways: for example, by helping to validate a result in the event of convergence, or by providing explanatory leads for erroneous results based on divergent findings. Taking this idea into legal territory, it could be argued that the opposition between explicable and non-explicable could be abandoned in favour of a global assessment of AI medical devices that takes into account the explainable approaches that have been used to develop it.

A third significant divergence concerns the value of informing the patient when AI is used. The legal literature devotes a great deal of attention to determining the conditions under which information should be provided (before or after the device is used) and the type of information that should be given to the person whose data is processed (general information, information about how the system works, specific information about the decision, etc.). The professionals interviewed did not volunteer any opinions on whether patients should be given information about the design or operating conditions of medical AI. It would seem that something is missing here. Could this "oversight" be attributed to the fact that the AI tool is in the development stage, some way from the ultimate stage of use? This would be tantamount to ignoring the fact that the interviewees have plenty to say about the need for doctors to be kept informed. It raises the question of how to involve patients in the design of the tools. Patients are largely absent from the multidisciplinary projects studied, even though patient groups or patient experts were sometimes included in research support committees. This is the case for the PRIMUS project, which is also distinctive in that it envisages a secondary use of the AI instrument as a means to improve communication with the patient by visualising results. On this point, should we be alert to the persistence of a form of medical paternalism or of a corporatist way of working (and of representing work) that demands technical autonomy? Conversely, might we ask whether there is anything different about AI systems compared with the measurement and visualisation tools already employed in healthcare without patients specifically being given information about how the technology works? By directly emphasising opacity as a source of difficulties that justify new obligations, digital law provides an opportunity for a fresh look at the question of the extent of a patient right to information. Should patients necessarily be informed of the technical characteristics and limitations of the tools chosen by their doctor to measure their biological constants or to visualise their internal lesions?

A fourth dissonance arises from the importance ascribed to information and explanation about the computer algorithm. These receive a great deal of attention in the legal literature, whereas the interviews and questionnaires reveal greater interest in the system's inputs and outputs. Data attract a great deal of attention, whether with respect to access, quality, annotation, structuring, biases and, finally, the transmission of information about them. Validation methods are the other major concern. It may not be inferred, however, that this supports the idea that explanation is of little importance, and that evidence of effectiveness is sufficient in medical practice (London, 2019). In fact, the problem of explainability is not eliminated, but rather appears to be posed differently. On the one hand, designers are adopting explainable or explicit technologies to complement those considered opaque. On the other hand, and more importantly, designers are seeking to make explicit the choices that governed the design and development of the medical device in which the algorithm is embedded, extending the question of "logic" to the medical reasoning behind the innovation. The matter of "explanation" thus takes different forms, as shown by the "Explanation of results" section of the multiple sclerosis clinical decision support system. This observation, which converges with certain studies in the human and social sciences (Winter and Carusi, 2023), raises the question of how to interpret the provisions of future regulation on AI.

CONCLUSION

In parallel with studies in the social sciences that explore the conditions of the development and deployment of medical AI (Vuarin and Steyer, 2023; Gillner, 2024; Anichini, 2023), it seems high time that legal experts took a more direct interest in the expectations and concerns of the users of automated decision support tools in order to obtain some grounding for their theories. Although it is not possible to draw any definitive conclusions from the data collected, we found several noteworthy disparities between the concerns expressed in the legal literature and those of medical professionals. The concerns expressed by the latter seem surprisingly similar to older legal analyses concerning automated tools that predate deep learning. In the 1981 decision by CNIL (the French data protection agency) on the GAMIN software, a decision support tool designed by the Ministry of Health to detect children at risk of disability and hence to

prioritise socio-medical monitoring for them, it was noted that a critical assessment of its operation was based primarily on the data used and the errors in its results (Huttner, 2022, p. 3). Although GAMIN's algorithm worked according to explicit rules, evaluation could already be carried out from the system's inputs and outputs. Why should this approach not be considered valid when the data processing rules or criteria are not known? Precisely because the path that leads from the inputs to the outputs generates uncertainty about the accuracy of the correlations made and that this is likely to affect confidence in the validation tests.

The transparency requirement introduced by the AI Act might increase misunderstandings if it prevented the deployment of AI deemed to be medically useful or, conversely, might offer common ground if it led to particular attention being paid to issues around the provision of information on data and performance. The AI Act contains a number of provisions that could generate agreement, providing reassurance that the new law is capable of encouraging virtuous practices rather than spawning new and unproductive demands. Moving away from the GDPR's focus on the "underlying logic" of the system, Article 13 links the obligation of transparency to "information that is relevant in explaining its results" and to "specifications of input data, or any other relevant information concerning training, validation and test datasets". This provision opens up the possibility of an interpretation that encourages the greatest possible effort to disclose and explain, using vocabulary that is highly consistent with the language of the professionals interviewed. If the effectiveness of a law depends on its capacity to reflect practices, which presupposes being sufficiently in tune with them to act on them effectively, the AI Act offers some guarantees here. However, this will mean overcoming certain representations of AI (as a blind and dangerous machine; Romele, 2024) and of the law (as a constraint that is useless because inconsistent with reality). For example, it may seem paradoxical that the AI Act, which appears to be better attuned to the concerns of professionals, nevertheless generates a perception of "high risk" that renders the description "AI" less desirable, or even encourages a preference for the development of a tool whose learning capacity is frozen before it comes into use in order to avoid running foul of the law.

REFERENCES

ANICHINI G. (2023), Automatiser le choix du meilleur traitement : enjeux et défis d'outil algorithmique pour le pronostic de la sclérose en plaques, *Cahiers François Viète*, série III, n° 15, p. 81-115.

ANICHINI G., GEFFROY B. (2021), L'intelligence artificielle à l'épreuve des savoirs tacites. Analyse des pratiques d'utilisation d'un outil d'aide à la détection en radiologie, *Sciences sociales et santé*, vol. 39, n° 2, p. 43-69.

BENBOUZID B., CARDON D. (2018), Machines à prédire, *Réseaux*, vol. 5, n° 211, p. 9-33.

BENSAMOUN A. (2023), *To be or not to be…* transparent, *Dalloz IP/IT & communication*, 3 December 2023, [Online] available at: https://www. dalloz-actualite.fr/node/ito-be-or-not-be-transparenti-pour-un-principe-matriciel-de-transparence-dans-l-environnement-n (consulted on 15/07/2024).

BERNELIN M., DESMOULIN S. (2021), L'intelligibilité des algorithmes dans les systèmes d'aide à la décision médicale, *International Journal of Bioethics and Science Ethics*, vol. 32, p. 19-31.

BIANCALANA M. (2023), Artificial Intelligence in Oncology: The Wins, The Challenges, and How We Can Deliver on Personalized Cancer Care, *Journal of Clinical Pathways*, vol. 9, n° 5, p. 55-59.

BORRILLO D. (2023), "Intelligence artificielle et traitement des données sanitaires en France", *in* PATRONI GRIFFI A. (ed.), *Bioetica, diritti e intelligenza artificiale*, Milan, Mimesis, p. 437-447.

CASEY B., FARHANGI A., VOGL R. (2019), Rethinking Explainable Machines: the GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise, *Berkeley Technology Law Journal*, vol. 34, no. 1, p. 143-188.

CASTETS-RENARD C. (2018), Régulation des algorithmes et gouvernance du machine learning : vers une transparence et "explicabilité" des décisions algorithmiques ?, *Revue droit & affaires*, 15e éd., n° octobre, p. 32-48.

CASTETS-RENARD C. (2020), Comment construire une intelligence artificielle *Recueil Dalloz*, p. 225-228.

CCNE, CNPEN (Comité national pilote d'éthique du numérique; 2022), *Diagnostic médical et intelligence artificielle: enjeux éthiques*, Joint opinion no. 141 for the CCNE and no. 4 for the CNPEN.

CERASA A., CROWE B. (2024), Generative Artificial Intelligence in Neurology: Opportunities and Risks, *European Journal of Neurology*, vol. 31, e1623, doi.org/10.1111/ene.16232

DESMOULIN-CANSELIER S., LE MÉTAYER D. (2018), Algorithmic Decision Systems in the Health and Justice Sectors: Certification and Explanations for Algorithms in European and French Law, *European Journal of Law and Technology*, vol. 9, no. 3, [Online] available at: https://ejlt.org/index.php/ejlt/article/view/626 (accessed 15/07/2024).

DUCHESNE S. (2000), "Pratique de l'entretien dit "non directif"*", in* BACHIR M. (dir), *Les méthodes au concret. Démarches, formes de l'expérience et terrains d'investigation en science politique,* Paris, PUF, p. 9-30.

EDWARDS L., VEALE M. (2017), Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for, *Duke Law & Technology Review,* vol. 16, n° 1, p. 18-84.

EON-JAGUIN F. (2022), Le médecin, véritable décideur et non simple auxiliaire de l'algorithme, *Dalloz IP/IT*, no. 1, p. 29-35.

EYNARD J. (2020), "Réflexions pour une intelligence artificielle digne de confiance", *in* DE GROVE-VALDEYRON N. (ed.), *Télémédecine et intelligence artificielle en santé : quels enjeux pour l'Union européenne et les Etats membres*, Toulouse, Presses de l'université Toulouse-Capitole, p. 165-192.

GAGLIO G., LOUTE A. (2023), L'émergence d'enjeux éthiques lors d'expérimentations de logiciels d'intelligence artificielle. Le cas de la radiologie, *Réseaux*, vol. 4, n° 240, p. 145-178.

GILLNER S. (2024), We're Implementing AI Now, so Why not Ask Us What to Do? - How AI Providers Perceive and Navigate the Spread of Diagnostic AI in Complex Healthcare Systems, *Social Science & Medicine*, vol. 340, p. 116-442.

GOODMAN B., FLAXMAN S. (2017), European Union Regulations on Algorithmic Decision-making and a "Right to Explanation", *AI Magazine*, vol. 38, no. 3, p. 50-57.

GROTE T., BERENS P. (2020), On the Ethics of Algorithmic Decision-making in Healthcare, *Journal of Medical Ethics*, vol. 46, no. 3, p. 205-211.

G'SELL F. (2020), "Les décisions algorithmiques", *in* G'SELL F. (ed.), *Le* big data *et le droit*, Paris, Dalloz, p. 104-113.

HAS (2023), *Guide d'aide au choix des dispositifs médicaux numériques à usage professionnel à destination des professionnels et des établissements de santé*, [Online] available at: https://www.has-sante.fr/upload/docs/application/ pdf/2023-06/dispositif_medicaux_numerique_a_usage_professionnel_guide_ daide_au_choix.pdf (consulted on 05/11/2024).

HUTTNER L. (2022), "La décision de l'algorithme. Étude de droit privé sur les relations entre l'humain et la machine", doctoral thesis in law, University of Paris 1, Panthéon-Sorbonne.

JATON F. (2020), *The Constitution of Algorithms. Ground-Truthing, Programming, Formulating*, Cambridge, The MIT Press.

KAMINSKI M. E. (2019), The Right to Explanation, Explained, *Berkeley Technology Law Journal*, vol. 34, no. 1, p. 189-218.

KORICA M., MOLLOY E. (2010), Making Sense of Professional Identities: Stories of Medical Professionals and New Technologies, *Human Relations*, vol. 63, n° 12, p. 1879-1901.

LOMBI L., ROSSERO E. (2024), How Artificial Intelligence is Reshaping the Autonomy and Boundary Work of Radiologists. A Qualitative Study, *Sociology of Health & Illness*, vol. 46, n° 2, p. 200-218.

LONDON A. J. (2019), Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability, *The Hastings Center Report*, vol. 49, no. 1, p. 15-21.

MALGIERI G., 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, no. 4, p. 243-265.

MAZEAU L. (2018), Intelligence artificielle et responsabilité civile : le cas des logiciels d'aide à la décision en matière médicale, *Revue pratique de la prospective et de l'innovation*, n° 1, p. 38-43.

MIGNOT L., SCHULTZ É. (2022), Les innovations d'intelligence artificielle en radiologie à l'épreuve des régulations du système de santé, *Réseaux*, vol. 2, n° 232-233, p. 65-97.

MITTELSTADT B., ALLO P., TADDEO M., WACHTER S., FLORIDI L. (2016), The Ethics of Algorithms: Mapping the Debate, *Big Data & Society*, No. 3(2),[Online] available at: https://journals.sagepub.com/doi/10.1177/2053951716679679 (accessed 05/11/2024).

MITTELSTADT B. (2022), *L'impact de l'intelligence artificielle sur les relations patients médecins*, report submitted to the Council of Europe, [Online] available at: https://www.coe.int/fr/web/bioethics/report-impact-of-ai-on-the-doctor-patient-relationship (consulted on 15/07/2024).

MORLEY J., MACHADO C., BURR C., COWLS J., JOSHI I., TADDEO M., FLORIDI L. (2020), The Ethics of AI in Health Care: A Mapping Review, *Social Science & Medicine*, vol. 260, [Online] available at: https://doi.org/10.1016/j.socscimed.2020.113172 (accessed 05/11/2024).

MUEHLEMATTER U. J., DANIORE P., VOKINGER K. N. (2021), Approval of Artificial Intelligence and Machine Learning-based Medical Devices in the USA and Europe (2015-20): A Comparative Analysis, *The Lancet Digital Health*, vol. 3, no. 3, [Online] available at: https://www.thelancet.com/journals/ landig/article/ PIIS2589-7500(20)30292-2/fulltext (consulted on 05/11/2024).

NETTER E. (2020), What is the purpose of the principle of transparency in personal data law, *Dalloz IP/IT*, no. 11, p. 611-615.

WHO (2021), *Ethics and governance of artificial intelligence for health: WHO guidance*, report, [Online] Available at: https://www. who.int/en/publications/i/ item/9789240037403 (consulted on 15/07/2024).

PASQUALE F. (2016), *Black Box Society: The Secr*et al*gorithms that Control Money and Information*, Cambridge (Mass.), Harvard University Press.

POWERS T. M., GANASCIA J.-G. (2020), "The Ethics of the Ethics of AI", *in* DUBBER M. D., PASQUALE F., DAS S. (eds), *The Oxford Handbook of Ethics of AI*, New York, Oxford University Press, p. 26-51.

ROCHFELD J. (2018), L'encadrement des décisions prises par algorithme, *Dalloz IP/IT*, n° 9, p. 474.

ROMELE A. (2024), *Digital Habitus: A Critique of the Imaginaries of Artificial Intelligence*, New York, Routledge.

VÉRON P. (2023), La responsabilité médicale et hospitalière à l'épreuve des systèmes automatisés d'aide à la décision, *Revue de droit sanitaire et social*, vol. 10, n° 5, p. 899-913.

VUARIN L., STEYER V. (2023), Le principe d'explicabilité de l'IA et son application dans les organisations, *Réseaux*, vol. 4, n° 240, p. 179-210.

WACHTER S., MITTELSTADT B., 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, p. 76-99.

WACHTER S., MITTELSTADT B., RUSSEL C. (2018), Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, *Harvard Jornal of Law & Technology*, vol. 31, no. 2, p. 841-888.

WINTER P. D., CARUSI A. (2023), (De)troubling Transparency: Artificial Intelligence (AI) for Clinical Applications, *Medical Humanities,* vol. 49, n° 1, p. 17-26.

ZHANG Y., WENG Y., LUND J. (2022), Applications of Explainable Artificial Intelligence in Diagnosis and Surgery, *Diagnostics (Basel)*, vol. 12, no. 2, p. 237, [Online] available at: https://doi: 10.3390/diagnostics12020237 (accessed 05/11/2024).