cepej

European Commission for the Efficiency of Justice

Commission européenne pour l'efficacité de la justice

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

8 December 2020

CEPEJ(2020)15Rev

EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE (CEPEJ)

Possible introduction of a mechanism for certifying artificial intelligence tools and services in the sphere of justice and the judiciary:

Feasibility Study

In December 2018, the European Commission for the Efficiency of Justice (CEPEJ) adopted the Ethical Charter on the use of artificial intelligence in judicial systems and their environment. The CEPEJ Charter represents the first step in the CEPEJ's efforts to promote responsible use of artificial intelligence (AI) in European judicial systems, in accordance with the Council of Europe's values. Mindful of the need to support the implementation of the Charter, the CEPEJ Working Group on the Quality of Justice (CEPEJ-GT-QUAL) has explored the possibility of introducing a mechanism for certifying AI solutions in accordance with the principles of the Charter.

This feasibility study was drawn up under the supervision of CEPEJ-GT-QUAL by Mr Matthieu Quiniou (France), a scientific expert, barrister at the Paris Court of Appeal and lecturer/researcher at Paris 8 University.

As adopted at the 34th plenary meeting of the CEPEJ,
8 December 2020

Contents

Feasibility study on the possible introduction of a mechanism for certifying artificial intelligence tools and services

Introduction

Justice is not currently the preferred focus of companies that are innovating in the field of artificial intelligence. That much is clear from the Forbes 2019 classification of America's most promising artificial intelligence companies for example, which does not include a single company that could be categorised as operating in the legal or judicial spheres[1]. This can be put down to the specific nature of the law market, its complexity, the regulatory barriers blocking access to it and also its minor economic importance. According to International Monetary Fund estimates, the law market worldwide constitutes barely more than 1% of the world's annual GDP, accounting for around 1 trillion[2] of a total 80 trillion euros[3]. A French study gauging the economic importance of the law market described it as being equivalent to that of the air transport, advertising or drinks sectors[4]. And yet artificial intelligence in the legal sphere is an issue of primary importance for society when it comes to predictive policing or automation of judicial decisions.

A number of "LegalTech" companies innovating in the legal sector and some States are now offering services using artificial intelligence to improve and personalise the results of legal database searches and facilitate decision-making, notably for provision in a context of litigation. The most ambitious mechanisms go as far as calculating the risks of reoffending, although such tools have sometimes been discarded after being experimented with on ethical grounds or because they performed poorly, as was the case with the COMPAS system in the United States[5].

This study is based mainly on the CEPEJ European Ethical Charter on the use of artificial intelligence (hereinafter "the Charter"), as a direct follow-up to it, but also on the following works:

- Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, Council of Europe study, DGI(2017)12 prepared by the Committee of experts on internet intermediaries (MSI-NET);
- White Paper on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final.

This study also took note of the ongoing work of the ad hoc Committee on Artificial Intelligence (CAHAI), recently established by the Council of Europe's Committee of Ministers.

The in-depth study on the use of AI in judicial systems, notably AI applications processing judicial decisions and data, reproduced in Appendix I to the CEPEJ Charter[6], defines the main categories of artificial intelligence in the sphere of justice for illustrative purposes as follows:

- Advanced case-law search engines
- Online dispute resolution,
- Assistance in drafting deeds
- Analysis (predictive, scales),
- Categorisation of contracts according to different criteria and detection of divergent or incompatible contractual clauses,
- "Chatbots" to inform litigants or support them in their legal proceedings.

Two additional categories with strong ethics implications could also be discussed, namely algorithmic justice, which could be seen as comparable to the aforementioned category of "online dispute resolution", and tools for enhanced judges' decision-making, which could be linked to the "Analysis" category, still at an experimental stage chiefly geared to aid for determining damages and sanctions.

[1] https://www.forbes.com/sites/jilliandonfro/2019/09/17/ai-50-americas-most-promising-artificial-intelligence-companies/#27f6479e565c

[2] https://www.prnewswire.com/news-releases/legal-services-market-to-be-driven-by-globalization-and-reach-1-trillion-by-2021-the-business-research-company-300886604.html

[3] https://www.imf.org/external/pubs/ft/weo/2017/02/weodata/weorept.aspx?sy=2010&ey=2017&scsm=1&ssd=1&sort=country&ds=.&br=1&pr1.x=60&pr1.y=13&c=001%2C998&s=NGDPD&grp=1&a=1

[4] Day One and Bruno Deffains, Le poids économique du droit en France, 2015, 22 pages.

[5] Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, How We Analyzed the COMPAS Recidivism Algorithm, Propublica, 2016 https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[6] CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems and their environment, Appendix 1, page 17.

As regards certification, these technological solutions are sufficiently different in terms of purpose and the use made of artificial intelligence that it might be envisaged either to certify only one category of those uses or to establish several sub-categories of certification with differentiated procedures and criteria.
At the CEPEJ-GT-QUAL meeting on 18 June 2020, it was emphasised that the cases of AI use that could be assimilated to predictive justice had the strongest ramifications for fundamental rights and freedoms and should therefore be focused on within the present study.

In addition to this diversity in terms of technological progress, functionalities and uses of artificial intelligence applications in the judicial sphere, we should also take into consideration the fields of law and specific characteristics of courts by legal sector. It should firstly be pointed out that some fields of law are more sensitive than others, owing to the ramifications of judicial decisions for personal freedoms. Criminology and criminal law are generally identified as fields of law where the use of artificial intelligence is to be envisaged with the utmost caution. Furthermore, to take the specific case of France, decisions are handed down, depending on the case, by courts made up solely of professional judges, by lay auxiliary judges, by trade union representatives or by juries. Other judicial systems, often in the English-speaking world, make a point of allowing judges' dissenting opinions to be transcribed. This diversity of judges is a source of creativity and a means of building bridges between society and law. There are numerous reports highlighting the risk of case-law becoming ossified by artificial intelligence reproducing carbon-copies of positions definitively laid down by supreme courts (whose role is indeed to harmonise a constantly evolving law).

The main typology of artificial intelligence systems distinguishes between symbolic (or human-readable) systems and connectionist (or purely machine-readable) systems. Symbolic artificial intelligence systems hinge on rules laid down by their designer to enable machines to take decisions on the basis of a pre-defined model. This form of artificial intelligence has been used for many years in expert systems, and in the legal sphere it is chiefly used for chatbots and also for constraint programming, for example using tree structures to automatically generate complex contracts from simple questions. Connectionist artificial intelligence systems are probabilistic and operate through induction and iteration in order to look at data and deduce characteristics that might become indicators and be used to interpret data sets. Training connectionist artificial intelligence systems requires both big data and sufficient computing power to process the data within a limited time. Connectionist artificial intelligence systems are the kind that would be used, for example, to define a strategy for litigation or set a sanction on the basis of a substantial collection of case-law.

Fundamental rights and freedoms appertain to human ontology, which differs from the functional approach of machines, even so-called intelligent ones. Unlike inanimate objects or machines with pre-programmed or strictly circumscribed behaviour, it is precisely the ability of artificial intelligence to go beyond the statistical model proposed and its constant improvement through successive iterations that define it.

Certification of connectionist artificial intelligence may be carried out at several levels: at the level of the learning model (supervised or not, scientific validity of the protocol and learning bias), at the level of the data used (validity of the data item, securing of the database and exclusion of certain sensitive data) and at the level of the results (adverse impact on fundamental rights and freedoms) .

The most evolved forms of artificial intelligence - systems that are connectionist with unsupervised learning and adaptative (non-deterministic) - process data on a contextualised basis, making discrimination breaching fundamental rights and freedoms more difficult to spot for humans not assisted by meta-intelligence that is specifically trained for the task. Furthermore, the constantly evolving nature of artificial intelligence necessitates continuous certification monitoring, possible for example by adapting the model used for indexing robots and by creating plugins.

Besides these considerations linked to the categories of artificial intelligence and its state of technological advancement in terms of contextual learning, we should distinguish between AI applications in the justice sphere according to their functions and purposes.

The response of CEN-CENELEC to the White Paper in June 2020[7] (COM(2020) 65 final) stressed the importance of considering specific applications and sector-specific issues where standards and certification of artificial intelligence systems are concerned[8]. Noting the lack of established standards in the field of artificial intelligence, CEN-CENELEC proposes in its recommendations that inspiration be drawn from the EU Ecolabel,

[7] CEN-CENELEC response to the EC White Paper on AI, June 2020.

[8] *Ibid*. p.6 "Furthermore, depending on the specific application various aspects (safety, fairness, privacy, security) have different relevance which must be considered by such a labelling scheme".

the certification scheme for cybersecurity now being finalised and national labelling schemes in Denmark and Malta[9].

The questions focused on as a priority in the present study will relate to the typology and challenges of certification and labels (I) and the objectives of CEPEJ certification (II). The study will also set out the issues linked to certification deployment, in terms of certification authorities (III), governance structure (IV), as well as the risks and opportunities entailed in such certification by the CEPEJ (V) and the issues of responsibilities linked to certification deployment (VI). The future European Union regulations on artificial intelligence will be a further consideration for the study (VII) and a schedule and roadmap will be suggested (VIII).

I. State of play, typology and challenges of standards, certification and labelling

The ethical certification of artificial intelligence appears to be an important issue in strategic areas in which algorithmic black boxes are inadmissible in a democratic society. European companies in strategic sectors are in favour of reliable and explainable certified artificial intelligence, often in contrast to the big American IT companies (GAFAM)[10]. European companies' need for certification is apparent, for example, in a statement by Groupe Thales chairman Patrice Caine, who said: "Artificial intelligence will soon be at the centre of all of our daily lives. Our customers in the aerospace, space, ground transportation, defence and security sectors are in charge of systems that are critical to the security of our societies, so it is vitally important to ensure that AI works as intended, to explain why it behaves in a given way, and to verify its use. That is why Thales is committed to making AI trustable, explainable, certifiable and ethical"[11].

I.1 Implications of the choice between mandatory and optional certification

Generally speaking, the main aim of technical standards is to cater for market needs, particularly in terms of compatibility, and certification also acts as an incitement to encourage the actors involved to adopt a certain conduct. For them, standards and certification systems can determine the launch of products on the market, how products or services are presented and also the criteria to meet when responding to public and sometimes private calls for tender.

In principle, both certification and compliance with technical standards for products or services are on a voluntary basis. That said, there may be regulations making certain technical standards mandatory, such as the European Union directives on the CE marking of certain products.

One purpose of technical standards and certification systems is to enable consumers to easily determine product or service quality in market comparisons, which in principle promotes competition. However, when such technical standards and certification are difficult to attain, this may act as a brake on competition by making it difficult for newcomers to gain access to the market, particularly in the case of mandatory or semi-mandatory standards or certification.

It should also be noted that mandatory or quasi-mandatory labelling preventing services from using artificial intelligence black boxes that cannot be interpreted or fail to comply with ethical principles could very significantly impact the ability of large tech companies to market their services in certain regions, unless they rethink the way they do things. Labelling or certification of this kind could set objective criteria for protecting users of such services and avoid diplomatic rows, such as the one in 2020 between the United States and the Chinese company TikTok[12].

I.2 Differences in the certification of symbolic and connectionist artificial intelligence systems

Symbolic artificial intelligence is based exclusively on models implemented by their designer; this type of artificial intelligence is deterministic. Symbolic artificial intelligence hinges on decision trees that may be fully interpreted and audited. A criterion that is discriminatory could be easily identified and modified, the certification

[9] *Ibid.* p.6 "Consider existing labelling schemes in Europe and the correspondent standards as inspiration, e.g. EU Ecolabel, upcoming certification scheme for Cybersecurity (Cybersecurity Act) and national AI labelling schemes (e.g. Denmark, Malta, etc).
8.2 To build a trustworthy and reliable label, standards are needed. Standards which currently do not yet exist. To first promote the development and acceptance of these standards, before introducing a labelling scheme. After introduction of a labelling scheme it is important to keep evaluating its effects."

[10] Google, Apple, Facebook, Amazon and Microsoft.

[11] "Thales illustre le rôle crucial de l'Intelligence Artificielle dans les moments décisifs" [*Thales demonstrates the critical role of AI in decisive moments*], Thales website, 17 January 2019 https://www.thalesgroup.com/fr/group/press-release/thales-illustre-le-role-crucial-lintelligence-artificielle-moments-decisifs

[12] https://www.reuters.com/article/us-usa-tiktok-china-pompeo-idUSKBN2480DF

of this type of artificial intelligence is fairly straightforward to implement and this can be done using database auditing criteria, criteria formulation and the decision tree.

In contrast, standardising and certifying connectionist artificial intelligence, especially systems functioning with unsupervised learning, is complex and still at the outline stage. Currently, the main challenge is gaining a precise understanding of the steps in the machine's reasoning without restricting its scope for producing original and pertinent interpretations. Some authors have suggested two distinct methods for rendering artificial intelligence interpretable: either to make the machine less complex or to apply a method which analyses the model after the training phase[13].

At present, there is no system for directly interpreting a connectionist artificial intelligence model using unsupervised learning. Researchers generally distinguish between explainability (commonly abbreviated to XAI, for "explainable artificial intelligence"), which is the ability to understand the general effects of the variables (the why,) and interpretability, which is the ability to gauge the importance of each variable in the result (the how) of an algorithm.

Methods such as LIME (*Local Interpretable Model-agnostic Explanations*) are used to facilitate understanding of the results coming out of neural network black boxes[14]. Evaluations of the quality of interpretation have already been put forward, and some authors have pointed to ethical problems arising from "producing explanations that are more persuasive than transparent"[15].

Studies carried out like those under the *TuringBox* project[16] undertaken by researchers from the Massachusetts Institute of Technology (MIT) may also be useful in facilitating work to interpret artificial intelligence systems in the legal sphere. This project enables individuals without specific IT expertise to examine an artificial intelligence system by stimulating machine behaviour and analysing the metrics. One of the main plus-points of this tool is its community of contributors which makes it possible to share the analyses carried out.

Specifically in the judicial field where artificial intelligence tools provide aid for decision-making for judges, if, for example, the machine proposes a detention measure rather than a security measure, the defendant and society are entitled to demand an explanation for that choice if it guides an official decision of justice. This necessity follows on from Article 5 - "right to liberty and security of person" of the European Convention for the protection of human rights and fundamental freedoms (hereinafter "the ECHR") entailing the obligation of reasoning of decisions and the requirement of non-arbitrariness[17]. The reasoning of decisions of justice demands a certain level of explainability. In the aforementioned example of an artificial intelligence system proposing a detention measure, depending on the type of artificial intelligence used (symbolic or connectionist, with or without supervised learning) and the artificial intelligence interpretation tools made available to the judge, there will be three possible scenarios: either the judge will be totally incapable of explaining the artificial intelligence decision; or the judge will be able to identify only one criterion that is irrational, irrelevant or not a sufficient basis for a legal argument; or the judge will be able to clearly explain the decision taken by the artificial intelligence application. In the first two cases, great care must be taken to avoid seeking to arrive at the same finding as the artificial intelligence system in terms of the sanction via another reasoning that would be explainable.

I.3 Certifying an evolving system

By nature, connectionist artificial intelligence is non-deterministic and capable of evolving. Only certain invariants such as data sources or the initial learning model (for those with supervised learning) can be certified on a lasting basis. The successive iterations of connectionist artificial intelligence render it more complex and, in principle, more relevant, and yet this increased complexity renders the explainability of the results more complex.

[13] See Christoph Molnar*, Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, lulu.com, 2020, 318 pages.

[14] See for example, Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Association for Computational Linguistics, 2016, 10 pages. Thomas Lin Pedersen, Michael Benesty, "lime: Local Interpretable Model-Agnostic Explanations": https://rdrr.io/cran/lime/ ; Zhang, Zhongheng et al. "Opening the black box of neural networks: methods for interpreting neural network models in clinical applications." *Annals of translational medicine* vol. 6,11 (2018): 216. doi:10.21037/atm.2018.05.32.

[15] See for example: Christophe Denis and Franck Varenne, "Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique", French Conference on Artificial Intelligence (CNIA) - Artificial Intelligence Platform (PFIA), July 2019, Toulouse, France. pp.60-68.

[16] https://turingbox.mit.edu/ and https://arxiv.org/abs/1803.07233

[17] European Court of Human Rights, Guide on Article 5 of the European Convention on Human Rights, Right to liberty and security, updated 30 April 2020, p.14.

For this type of non-deterministic artificial intelligence system, certification has to be carried out continually to be effective. Accordingly, it is possible to envisage continuous human monitoring for certified artificial intelligence systems, with the creation, for example, of a function of artificial intelligence officer within the companies or administrations proposing this type of certified artificial intelligence, along the lines of a data protection officer (hereinafter "DPO"), or possibly automated monitoring by a symbolic artificial intelligence application, in the form of a plugin, checking that the connectionist artificial intelligence it is monitoring still meets the criteria on which it was initially certified.

I.4 Comparison with certification of personal data processing systems

The certification of artificial intelligence hinges on analysis of the algorithm, the learning model and the quality of the data sources. The main aims of certification of personal data processing systems are to verify compliance with personal data regulations, namely the General Data Protection Regulation (hereinafter "the GDPR")[18] in the European Union, which primarily entails certification of procedures for informing and obtaining the consent of the individuals whose data are collected and processed and, residually, certification of the precautions and steps taken to secure those data.

As regards personal data since the GDPR's entry into force, the independent national data protection authorities (members of the G29), like the CNIL in France, have abandoned labelling linked to governance procedures for data protection, digital safe services labelling and audit procedure labelling, and now instead focus solely on a standard for certifying DPOs and certification bodies[19]. As things stand, these forms of certification can only be a source of inspiration if the aforementioned function of artificial intelligence officer is created. It should be noted that additional certification is almost certain to emerge in various European Union Member States, as Article 42 of the GDPR "Certification" encourages certification mechanisms and labelling for the purpose of demonstrating that processing operations comply with the GDPR.

There are also ISO standards in the area of personal data protection, notably ISO/IEC 27701. This international standard heralded by the ISO as the first international standard for managing privacy information[20] was published in 2019. It reflects the state of the art as regards protection of privacy and, among other things, covers the requirements for creating, implementing and improving privacy information management systems (PIMSs)[21] and security issues. This standard may be seen as a prerequisite for a component of certification for an artificial intelligence system. It should be noted however that this ISO/IEC 27701 standard is not GDPR certification within the meaning of Article 42 of the GDPR and forms part of what has been described as a "grey zone" of standardisation prior to regulatory authorities placing certification on a formal footing[22].

I.5 CE marking of artificial intelligence solutions seen in context

CE marking is governed by Regulation (EC) no. 765/2008 in terms of its definition, format and general principles. CE marking is not a certification mechanism but a visible commitment on the part of a manufacturer to comply with European legislation. The European Commission's Blue Guide states: "By affixing the CE Marking the manufacturer declares on his sole responsibility that the product conforms to all applicable Union legislative requirements, and that the appropriate conformity assessment procedures have been successfully completed"[23]. CE marking is mandatory only for certain products, referred to in European directives. There is an implicit suggestion in the European Commission White Paper on artificial intelligence that a directive classifying artificial intelligence systems as products or services requiring CE marking may be an avenue to be explored[24].

Because the sector in which they operate brings them under one of these directives, some artificial intelligence applications are subject to assessments to qualify for CE marking, notably medical devices using artificial intelligence. In this context, for example in France, the National Health Authority has rolled out a "Draft evaluation grid for medical devices using artificial intelligence"[25], with a view to medical devices being included

[18] 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).

[19] For example in France for the CNIL, AFNOR or Bureau Veritas.

[20] https://www.iso.org/news/ref2419.html

[21] https://www.iso.org/standard/71670.html

[22] Eric Lachaud, "ISO/IEC 27701: Threats and Opportunities for GDPR Certification", 2020, 24 pages.

[23] European Commission Notice, The 'Blue Guide' on the implementation of EU products rules 2016 (Text with EEA relevance) (2016/C 272/01).

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

[25] https://www.has-sante.fr/jcms/p_3118247/fr/projet-de-grille-d-analyse-pour-l-evaluation-de-dispositifs-medicaux-avec-intelligence-artificielle

in the list of products and services that may be reimbursed. According to the National Health Authority, this assessment should be carried out after CE marking, which may explain why its assessment is organised by the National Health Authority downstream. That said, a concomitant or prior assessment might be envisaged, as this issue appears to be more of a consideration upstream of CE marking and an application for inclusion on the list of reimbursable products.

CE marking therefore seems to be of little relevance to what might be devised in terms of artificial intelligence certification, and the European Union does not appear to have embarked upon in-depth discussion of a directive that could create an obligation of CE marking specific to artificial intelligence systems and even less so to artificial intelligence in the legal sphere, an area where there is not a single product or service coming under a European directive linked to CE marking.

I.6 The Ecolabel for artificial intelligence solutions seen in context

The European Ecolabel was created by Council Regulation (EEC) No 880/92 of 23 March 1992 on a Community eco-label award scheme and then revised by Regulation (EC) No 1980/2000 of the European Parliament and of the Council of 17 July 2000 on a revised Community eco-label award scheme. This voluntary labelling scheme makes it possible to attest to the environmental quality of a product during its life cycle through compliance with criteria specific to each product group[26]. To obtain the Ecolabel, a product must be subjected to a procedure for certification by a certifying authority, and the applicant must pay the processing costs and an annual fee thereafter[27].

The main point of convergence between the Ecolabel and the certification of artificial intelligence in the judicial sphere, whose feasibility is discussed in the present study, is their societal objective and their universalist aspirations.

The Ecolabel strategy is based mainly on the comparison and evaluation of the products best meeting the environmental objectives[28] and takes account of technical progress with a view to regularly updating the criteria[29]. A system of certification for the ethical aspects of artificial intelligence could draw on this comparative and evolutive approach.

I.7 National standards being devised for artificial intelligence: Malta and Denmark

Malta embarked upon in-depth thinking on the regulation and certification of artificial intelligence, with the production of a consultation document on the ethical aspects of artificial intelligence[30] in August 2019, followed by a strategic document in October 2019 aimed at making Malta the *"Ultimate AI Launchpad"*[31].

The purpose of the consultation on the ethical aspects of artificial intelligence is to arrive at a code of ethics for artificial intelligence paving the way for respect of fundamental rights, among other things. This code of ethics, yet to be finalised, has some similarities with the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems but its scope is not limited to the legal and judicial spheres and the approach taken is intended to fit with Malta's strategy to be competitive in this sector. The final section of the consultation document focuses on the certification of artificial intelligence without going into any real detail, the stated aim being to create trust and transparency between users, consumers and stakeholders[32].

The strategy document proposes undertaking AI-related pilot projects in government and authorities in various sectors (traffic management, education, health care, customer service, tourism). The justice sphere has not been included in the areas listed for experimentation[33]. The sole reference to artificial intelligence in the justice sphere in this document is to be seen in a survey finding that only half of the respondents are comfortable with

[26] Article 6, Regulation (EC) No 1980/2000 of the European Parliament and of the Council of 17 July 2000 on a revised Community eco-label award scheme.

[27] *Ibid.*, Article 12.

[28] *Ibid.* Recital 6: "It is necessary to explain to consumers that the eco-label represents those products which have the potential to reduce certain negative environmental impacts, as compared with other products in the same product group, without prejudice to regulatory requirements applicable to products at a Community or a national level".

[29] *Ibid.* Article 4.4 and 16.

[30] Parliamentary Secretariat For Financial Services Digital Economy and Innovation, Office of the Prime Minister, Malta: Towards Trustworthy AI: Malta's Ethical AI Framework For Public Consultation, August 2019, 36 pages.

[31] Parliamentary Secretariat For Financial Services Digital Economy and Innovation, Office of the Prime Minister, Malta : the Ultimate AI Launchpad, October 2019, 57 pages.

[32] Parliamentary Secretariat For Financial Services Digital Economy and Innovation, Office of the Prime Minister, Malta: Towards Trustworthy AI: Malta's Ethical AI Framework For Public Consultation, August 2019, pp. 33-34.

[33] [33]Parliamentary Secretariat For Financial Services Digital Economy and Innovation, Office of the Prime Minister, Malta : the Ultimate AI Launchpad, October 2019, p.2.

AI in the area of justice, compared to 80% in the area of transportation[34], which appears to have been the reason why it was excluded from the pilot experiments.

Denmark is another country that has been actively looking at the labelling of artificial intelligence based on ethical criteria. A national strategy for artificial intelligence was set out in a government document published in March 2019, geared to creating a toolbox for companies and a data ethics label for artificial intelligence[35]. The forthcoming creation of labelling for IT security and responsible data use was announced at the end of 2019 by the Danish Ministry of Industry[36].

These national labelling strategies, welcomed by the European Union[37] and the OECD[38], are still at the design stage where the ethical certification of artificial intelligence is concerned and do not provide any pointers for a working basis or points of comparison for the present study. That said, these national efforts show that the Council of Europe's member States are keen on having certification systems in this sensitive area.

I.8 The work of the IEEE to set standards for ethics in the artificial intelligence field

The Institute of Electrical and Electronics Engineers (IEEE) is a non-profit organisation formed under American law which is highly active in developing electronics- and IT-related technical standards and describes itself as the world's largest technical professional organisation for the advancement of technology.

Beyond its technical standard-setting work, the IEEE is already at an advanced stage in work aimed at standardising "the future of ethics for autonomous and Intelligent systems", which takes the form of a series of projects numbered P7000™ ,. The projects are sequentially numbered up to 14 (skipping no. 13):

- IEEE P7000™: Model Process for Addressing Ethical Concerns During System Design
- IEEE P7001™: Transparency of Autonomous Systems
- IEEE P7002™: Data Privacy Process
- IEEE P7003™: Algorithmic Bias Considerations
- IEEE P7004™: Standard on Child and Student Data Governance
- IEEE P7005™: Standard on Employer Data Governance
- IEEE P7006™: Standard on Personal Data AI Agent Working Group
- IEEE P7007™: Ontological Standard for Ethically driven Robotics and Automation Systems
- IEEE P7008™: Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
- IEEE P7009™: Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- IEEE 7010™-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
- IEEE P7011™: Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources
- IEEE P7012™: Standard for Machine Readable Personal Privacy Terms
- IEEE P7014™: Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems

Of these ongoing projects, only *IEEE 7010™-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being* has produced a standard to date.

In the list of standardisation projects above, only P7000 to P7003 are truly relevant to the present study on the certification of artificial intelligence in the judicial sphere. An outline of the P7000 draft standard can already be ordered on the IEEE's site but no outline documents have been published yet for projects P7001 to 7003. These P700X draft standards being devised by the IEEE do not directly relate to the judicial sphere for the time being and have more to do with ethical conduct or well-being. The questions of transparency (P7001™) and algorithmic bias (P7003™) are key and will have to be taken into account once these IEEE standards have actually been drawn up, potentially providing pointers for refining the indicators useful for implementing Principle no. 4 of the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems and certification criteria (*see II below*).

[34] *Ibid.*, p. 38.
[35] Danish Government, National Strategy for Artificial Intelligence, March 2019, 74 pages: https://eng.em.dk/media/13081/305755-gb-version_4k.pdf
[36] https://eng.em.dk/news/2019/oktober/new-seal-for-it-security-and-responsible-data-use-is-in-its-way/
[37] White Paper on Artificial Intelligence, A European approach to excellence and trust, COM(2020) 65 final, p. 13.
[38] https://oecd.ai/wonk/an-independent-council-and-seal-of-approval-among-denmarks-measures-to-promote-the-ethical-use-of-data

I.9 Integration of labelling into the work of the Council of Europe (European Pharmacopoeia and Cultural Routes Label)

The European Pharmacopoeia, set up under the auspices of the Council of Europe by its member States by a convention in 1964[39], has gradually opened up to standardisation bodies. The Pharmacopoeia is run by the Council of Europe via its European Directorate for the Quality of Medicines and HealthCare (EDQM) which currently employs some 400 people of 25 different nationalities. The EDQM has signed a memorandum of understanding with the European Committee for Standardisation (CEN) concerning medical devices[40].

The factsheet on the EDQM says that the European Pharmacopoeia is the "official reference work used by professionals involved in the manufacture and control of medicines. Its objective is to define legally binding quality requirements for medicines and their ingredients"[41]. Compliance with the quality requirements of the European Pharmacopoeia (Ph. Eur.) is a prerequisite for medicines to be authorised for the markets of the 39 member States. The binding nature of this certification provides a solid foundation for the recommendations and criteria defined by the EDQM.

The reason for the EDQM operating within the Council of Europe is that, according to its mission statement, it "contributes to the basic human right of access to good quality medicines and healthcare, and promotes and protects human and animal health". The aim of certification discussed in the present study seems even more closely geared to the Council of Europe's human rights protection missions in seeking to adapt them to the digital universe, and particularly to artificial intelligence systems, in the judicial sphere.

While gained in an area that is very different from the legal and judicial spheres, the European Pharmacopoeia's experience may be highly instructive for devising binding certification prior to products coming onto the market and structuring a certification system from an institutional and organisational viewpoint (*see section IV below*) and in relation to standardisation bodies and the European Union.

The Council of Europe also runs the Cultural Routes labelling scheme. A look at the experience acquired could be useful for devising and deploying Council of Europe certification for artificial intelligence, from an organisational point of view (*see section IV below*) whether for defining the evaluation cycles or setting up a certifying body. The Council of Europe teams involved in implementing the Cultural Routes label could be asked to provide insight into the specific issues of a labelling system on the scale of the Council of Europe.

The experience gained with the Cultural Routes labelling could prove to be beneficial for identifying points of convergence and possible cooperation with other international institutions active in cultural labelling which have also carried out substantive work on artificial intelligence, especially the United Nations Educational, Scientific and Cultural Organisation (UNESCO) and, to a certain extent, the Organisation for Economic Co-operation and Development (OECD).

I.10 GDPR-style impact assessment and certification

Impact assessments carried out by companies have been made systematic by the GDPR, which imposes this type of documentation in its Articles 35 and 36: where personal data processing "is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data"[42].

An impact assessment for artificial intelligence systems could be envisaged along the lines of the GDPR model, taking account of not only the consequences for personal data protection but also and above all the societal and economic impact of inexplicable or biased decisions taken by artificial intelligence.

[39] Convention on the Elaboration of a European Pharmacopoeia, 1964, Strasbourg, 22.VII.1964.

[40] https://www.edqm.eu/en/History-1964-1997-1562.html

[41] https://www.edqm.eu/sites/default/files/medias/fichiers/Factsheets/factsheet_pheur_reference_standards_july_2020.pdf

[42] Article 35, 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).

I.11 Contextualised artificial intelligence sandboxes and certification

The term "sandbox" is commonly used in the IT field in connection with the testing of programmes before they are rolled out, particularly to spot bugs and ensure good launch conditions. This is chiefly an IT security practice allowing a programme to be run in a restricted environment[43].

This practice has recently been transposed to the legal sphere, with regulatory sandboxes, initially used in the area of financial regulation. The pioneering project in this field, *Project Innovate*, is a programme run since 2014 by a British financial regulation body, the *Financial Conduct Authority* (FCA). According to the FCA, it allows businesses to test innovative propositions in the market, with real consumers. Taking this approach makes it possible to test a product or service in a controlled environment, reduce time-to-market, help identify appropriate consumer protection safeguards to build into new products and services before they are put on the market and secure better access to project finance[44]. The practice has become widespread since 2014 and has a strong following throughout the FinTech sector and also in artificial intelligence[45], notably in France[46] or Finland[47].

With a view to Council of Europe certification of artificial intelligence in the judicial sphere, one benefit of a sandbox would be to make the actors involved aware of the challenges of human rights by design and ethics. A sandbox would provide a means of giving them guidance over a prolonged period as they get to grips with the content of the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems and certification criteria.

A sandbox would also make it possible to continuously refine and improve the certification criteria to closely match practice and developments on the technical side.

For those proposing artificial intelligence projects in the judicial sphere, in addition to the certification they may obtain during or at the end of sandbox testing, this support could help improve their product's chances of commercial success, as their customers (States, administrations, regulated law professionals etc) will be particularly attentive to compliance with regulations and the efforts made to achieve this. For projects pursued directly by States or administrations, a sandbox, and more generally certification in this area, would provide useful support in aligning them with supranational standards and reducing risks of the European Court of Human Rights ruling against them in a strategic area.

Nonetheless, since, as pointed out earlier, connectionist artificial intelligence applications are systems that evolve, the certification awarded, after sandbox testing for example, cannot be definitively acquired.

Since there is currently no artificial intelligence certification in the judicial sphere to serve as a source of inspiration, thoughts have been turned to the certification of artificial intelligence, and, more globally, invariants in the certification process make it possible to design such a tool. The experience gained by the Council of Europe in labelling and certification in connection with cultural routes and the European Pharmacopoeia may prove useful for structuring the certification envisaged in institutional terms. Moreover, this certification could be a means of disseminating and ensuring the effective application of the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems, through support and guidance methods such as sandboxes.

II. Formalising criteria and indicators for certification in line with the Charter

II.1 Methods for formalising indicators based on the Charter's principles

In order to faithfully transpose the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems in a certification scheme, we need to design operational, objective and verifiable indicators based on the five principles it sets out:

- Principle of respect for fundamental rights: ensure that the design and implementation of artificial intelligence tools and services are compatible with fundamental rights.

[43] VassilisPrevelakis and Diomidis Spinellis, "Sandboxing Application", Proceedings of the USENIX Technical Annual Conference, Freenix Track, June 2001, pp. 119–126 (2001).

[44] https://www.fca.org.uk/firms/innovation/regulatory-sandbox

[45] For more general information: Laura Delponte, "Study: European Artificial Intelligence (AI) leadership, the path for an integrated vision", European Parliament, IPOL, 2018, 48 pages.

[46] Cédric Villani, "For a meaningful artificial intelligence", French parliamentary mission, 2018, 233 pages.

[47] https://www.tekoalyaika.fi/en/reports/finland-leading-the-way-into-the-age-of-artificial-intelligence/3-eleven-key-actions-ushering-finland-into-the-age-of-artificial-intelligence/

- Principle of non-discrimination: specifically prevent the development or intensification of any discrimination between individuals or groups of individuals.
- Principle of data quality and security: with regard to the processing of judicial decisions and data, use certified sources and intangible data with models elaborated in a multi-disciplinary manner, in a secure technological environment.
- Principle of transparency, impartiality and fairness: make data processing methods accessible and understandable, authorise external audits.
- Principle "under user control": preclude a prescriptive approach and ensure that users are informed actors and in control of the choices made.

II.2 Indicators for Principle no. 1: respect for fundamental rights

> Principle no. 1 of the Charter: "Principle of respect for fundamental rights: ensure that the design and implementation of artificial intelligence tools and services are compatible with fundamental rights"

The first principle on respect for fundamental rights is the most specific and certainly the most complex to transpose into an artificial intelligence certification scheme. Among other things it hinges on the concept of *Human rights by design*, which serves as the key strategy line of the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems.

- Proportionate processing of personal data (privacy) and clear purposes

- Anonymisation of the parties and participants (physical individuals) and their counsels
  - Checking by consultation of data sets
- Absence of evaluation and classification of physical individuals or legal entities on the basis of judicial decisions
  - Checking by consultation of the interface
- Anonymisation of the name of the judge and the location of the court in decisions used for predictive justice (with the aim of avoiding *forum shopping*)
  - Checking by consultation of data sets
- Hermetic separation of artificial intelligence services having different purposes (such as dissociating the search engine service from the aid for decision-making service)
  - Checking of databases and data sources used by each system

- Right of access to the judge and the right to a fair trial

- Presence of clear information indicating, where applicable, that a report generated by an artificial intelligence system is not explainable. This aspect allows the judge to maintain full control in his decision-making and to use the reports not fully explicable, which can be generated by connectionist artificial intelligences, only knowingly.
  - Learning model and interface
- For aspects related to access to the judge, refer to indicators related to user control and more specifically to the possibility of AI opt-out for the defendant (II.6)

- Judges' independence in their decision-making process

- Safeguard against the profiling of judges
  - A/B testing checking on the search engine using different user accounts and different search histories
- Match between the criterion displayed and the actual pattern of classification of search results
  - Checking by auditing search results
- Transparency of weighting of criteria for multicriteria searches
  - Checking of the existence of explanatory information and auditing of search results
- Transparency of criteria used for searches by "relevance"
  - Checking of the existence of explanatory information and auditing of search results

- Ethics and Human rights by design

- Taking into account of fundamental rights and freedoms and the ECHR from the outset:
  - for symbolic artificial intelligence systems
    - Report presenting the decision trees explaining how fundamental rights and freedoms are to be taken into account
  - for connectionist artificial intelligence systems with supervised learning
    - Report presenting the training data and methods, explaining how fundamental rights and freedoms are to be taken into account

        - Assessment of the impact of the service using artificial intelligence in terms of fundamental rights and freedoms: Absence of prior checking (accountability of the actors)
        - Optional designation of an independent artificial intelligence officer where explainability is adequate and mandatory designation where explainability is inadequate
    - for connectionist artificial intelligence systems with unsupervised learning
        - Assessment of the impact of the service using artificial intelligence in terms of fundamental rights and freedoms: Absence of prior checking (accountability of the actors)
        - Compulsory designation of an independent artificial intelligence officer
- Installation of and continuous permission for access to a CEPEJ human rights conformity plugin (meta-AI/plugin which could be developed subsequently using the information obtained from impact assessments; such a plugin could potentially be used or adapted beyond the legal and judicial spheres).

II.3 Indicators for Principle no. 2: non-discrimination

> Principle no. 2 of the Charter: "Principle of non-discrimination: specifically prevent the development or intensification of any discrimination between individuals or groups of individuals"

The second principle is aimed at obviating discrimination that could arise in connection with sensitive data. By nature, artificial intelligence is dependent on the creation of more or less finely drawn categories, and therefore discrimination. Limiting discrimination means curbing artificial intelligence and removing sensitive data may create biases, which may be preferable or necessary. The anonymisation or selective exclusion of certain data is a means of limiting discrimination[48]. Some research, such as that carried out by researchers at Eindhoven University of Technology suggest solutions entailing the reprocessing of data in order to eliminate discriminatory effects[49] but, for the time being, these seem very onerous and difficult to deploy beyond an experimental environment. Some sensitive information, relating for example to someone's state of health may have the effect of diminishing their responsibility or constituting attenuating circumstances. Nevertheless, to effectively avoid discrimination on bases that would be contrary to human rights and not exacerbate inequality, these data should be comprehensively taken into account by including tags capable of indirectly producing the same discriminatory effects by cross-referencing (deducing a person's origin from their surname, deducing sexual orientation from contextual elements, deducing socio-economic background from a home address etc). It should also be noted that the use of artificial intelligence may well give rise to new heavily profiled and category-unrelated forms of discrimination that are less readily discernible, such as discrimination based on ethnic background.

This principle must be considered in parallel to the stipulations of Article 6 of Convention 108+ regarding special categories of data, and Articles 9 and 10 of the GDPR also applicable in many Council of Europe member countries, which authorise the processing of this type of data solely on condition that there are suitable safeguards complementing the obligations applicable to classic personal data.

Among the exceptions granted under Article 9 of the GDPR are processing necessary for the establishment, exercise or defence of legal claims and processing necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued.

Article 10 of the GDPR puts forward a distinction as to the entity carrying out the processing of personal data relating to criminal convictions and offences or related security measures, which may be processed only under the control of official authority and in exceptional cases and with appropriate safeguards from other entities.

Article 6 of Convention 108+ expressly refers to the risk of discrimination resulting from the processing of sensitive data, which is also emphasised in the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems.

[48] On this point, see for example: Salvatore Ruggieri, "Data Anonymity Meets Non-Discrimination", IEEE 13th International Conference on Data Mining Workshops (December 7-10, 2013), pp. 875–88.

[49] See for example Faisal Kamiran and Calder Toon, "Data Preprocessing Techniques for Classification Without Discrimination", Knowledge and Information Systems 33, no. 1 (December 3, 2011), pp. 1–33.

- Avoiding discrimination based on sensitive data

On principle:

- Elimination of the tags that could be linked to parties' sensitive data (home address, income, family situation, registered capital)
  - Checking by consultation of data sets
  - A/B testing using information and tags that could be linked to sensitive data by changing, where applicable, one of the following parameters in each test: surname, home address, income, family situation, registered capital, relevant specific contextual information
  - Form enabling users to submit detailed requests for the removal of tags that could be linked to sensitive data, with copy to the artificial intelligence officer, where applicable, and to the control authority (labelling entity).

By way of exception, for artificial intelligence systems under public authority control and for data relating to criminal convictions and offences or related security measures:

- Mechanisms guaranteeing that the processing is necessary and proportionate
  - Impact assessment by the processing body explaining the safeguards implemented
  - Designation of an artificial intelligence officer

II.4 Indicators for Principle no. 3: data quality and security

> Principle no. 3 of the Charter: "Principle of quality and security: with regard to the processing of judicial decisions and data, use certified sources and intangible data with models elaborated in a multi-disciplinary manner, in a secure technological environment"

The previously mentioned ISO/IEC 27001 standard (see section I.5 above) represents, in 2020, the benchmark for IT and data security. Standard ISO/IEC 27001 provides for a hundred or so security measures designed to control the solidity of a system in terms of cybersecurity. Compliance with this technical standard might be stipulated as a prerequisite for this certification scheme. In the area of cybersecurity, using an ISO/IEC technical standard, designed chiefly by companies in the sector, does not appear to pose any structural difficulties as those companies seek to apply the highest standards in order to fulfil their obligation of result for their customers and comply with the relevant laws and binding regulations. When certification within the meaning of Article 42 of the GDPR is adopted in the area of cybersecurity, it could be seen as a replacement for the ISO/IEC standard.

Ensuring the quality of data and that it does not undergo semantic alteration[50] up until integration in the data set provided for an artificial intelligence system is an issue of fundamental importance, commonly designated by the acronyms ALCOA[51] and ALCOA+[52], which stipulate that data must be attributable, legible, contemporaneous, original, accurate, complete, consistent, enduring, available.

One possible method of ensuring the integrity of the "data based on judicial decisions" referred to in the third principle of the Charter might be to create a system enabling the competent authorities in member States to blockchain digitised judicial decisions. However, certification schemes, particularly when mandatory, can in some cases create indirect obstacles to applications coming onto the market.

This third principle also evokes the multidisciplinary aspect of designing or at the very least evaluating artificial intelligence models. Such an objective might be achieved by setting up a group of experts made up of law professionals, researchers and academics from disciplines including law, ICT science, mathematics, IT, economics or sociology. This group could provide the scientific structure for a mechanism for giving support and guidance to promoters of artificial intelligence solutions in the legal and judicial sphere, particularly within the framework of sandboxes (see section I.11 above).

II.5 Indicators for Principle no. 4: transparency, impartiality and fairness

> Principle no. 4 of the Charter: “Principle of transparency, impartiality and fairness: make data processing methods accessible and understandable, authorise external audits.”

[50]Semantic alteration does not include, for example, the exclusion of sensitive data, such as those whose removal has previously been envisaged (see. II.3).

[51] Acronym standing for: Attributable; Legible; Contemporaneous; Original; Accurate.

[52] The "+" in ALCOA+ stands for: Complete ; Consistent ; Enduring ; Available.

- Access to source code for certification

The references cited in the Charter put this principle into perspective, particularly the Council of Europe's MSI-Net study, the Villani report and the report by the House of Lords, which note with some resignation the unlikelihood of securing the provision of entire algorithms or their source codes and merely entertain the possibility of obtaining key subsets of information.[53] In the case of optional certification (see I.1), which provides for full disclosure of the source code as a condition for receiving certification, the companies with the most advanced algorithms may not wish to relinquish their competitive advantage even in return for certification that is likely to reassure their customers.

In the case of mandatory certification, imposing obligatory source code access appears to be perfectly feasible but should be implemented with the utmost respect for trade secrets.

It would therefore be possible to consider the confidential transmission of entire source codes to the committee awarding the certification or to the artificial intelligence officer, possibly after *intuitu personae* approval by the company in question.

Furthermore, state-controlled artificial intelligence applications that are designed in-house by public authorities should be open source, in the same way as legal texts are accessible to citizens, with the selection of artificial intelligence designed by service providers and used by States conditional on it being open source.

For some types of artificial intelligence, source code access does not always suffice to ensure that it is fully explainable.

- FAT-ML: artificial intelligence that is fair, accountable and transparent

In the study of the behaviour of algorithms, there is some consensus for focusing on the fair, responsible and transparent character of artificial intelligence; this approach is called FAT-ML (Fairness, Accountability and Transparency in Machine Learning)[54].

- Fairness

The issue of the fairness of artificial intelligence intersects with that of non-discrimination referred to in Principle no. 2 of the Charter (see II.3).

- Accountability for artificial intelligence

The criterion of accountability refers to the possibility of holding a natural or legal person liable in the event of artificial intelligence malfunctioning or causing damage. Certification could potentially mean the legal entity acting in good faith could be at least partly exonerated from liability in the event of a problem arising from certified artificial intelligence. The aforementioned sandbox method is designed to limit the damage, and therefore the liability that may be incurred, through initial deployment of the application in a controlled and restricted environment during a testing phase.

- Transparency

Beyond access to code, transparency refers to the possibility of explaining the decisions of artificial intelligence systems. This aspect of artificial intelligence is vital and has already been widely taken on board in the indicators proposed for the first principle as a prerequisite of any ethical by design and human rights by design system, which are distinguished according to the type of artificial intelligence — symbolic or connectionist (see II.2).

[53] The CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems, page 11, footnote 3: "Of interest in this connection is the suggestion made on page 38 of the Council of Europe's MSI-NET study on 'Algorithms and Human Rights': 'The provision of entire algorithms or the underlying software code to the public is an unlikely solution in this context, as private companies regard their algorithm as key proprietary software that is protected. However, there may be a possibility of demanding that key subsets of information about the algorithms be provided to the public, for example which variables are in use, which goals the algorithms are being optimised for, the training data and average values and standard deviations of the results produced, or the amount and type of data being processed by the algorithm.'" Or even the suggestions appearing on page 117 of the aforementioned "AI for Humanity" report drafted by Mr Cédric Villani, a member of the French National Assembly, as part of a mission assigned to him by the Prime Minister of the French Republic: "The auditors may be satisfied with simply checking the fairness and equity of a programme (doing only what is required of them), by submitting a variety of false input data, for example, or by creating a large quantity of system user profiles according to precise guidelines." In addition, there are also the statements in the report by the House of Lords, "AI in the UK: ready, willing and able?", paragraphs 92, 96-99.

[54] For more information, see the website: https://www.fatml.org/

Symbolic artificial intelligence systems are deterministic, rigid and explainable because they model human thought processes with representations. Provision of the model is therefore all that is required to audit this type of artificial intelligence.

Connectionist artificial intelligence systems may become harder to explain as they evolve and complexify. For this type of artificial intelligence, a system of continuous monitoring conducted by an artificial intelligence officer or an automatic verification plugin should be considered with a view to explainability. When a full explanation cannot be provided, it must at least be possible for users accessing the result supplied by artificial intelligence to be immediately aware that this is the case (see I.2 and I.3).

With regard to transparency, the IEEE P7001™ Transparency of Autonomous Systems standard which is currently under development should make it possible to integrate technical specifications with the aim of reinforcing the protection of the principle of transparency.

- Identifying artificial intelligence systems and their actions

The question of identification is crucial to ensuring the transparency and responsibility of actors involved in the sector. Just as a user is indirectly identifiable when browsing and communicating, it would be useful if this were also the case for an artificial intelligence system, which has a certain autonomy in its actions. An AI system should be explicitly identified as such on the network (possibly with the indication of its category, connectionist or symbolic). Today, users can be indirectly identified on a network via their hardware's MAC address or to some extent on the Internet via the IP address (if static) of the network-connected device. Enhanced user identification tools are used for processes requiring secure identification (KYC systems, electronic signatures, etc.).

Identifying artificial intelligence systems means that the information-gathering operations they perform can be seen (they leave a readable trace) and their actions can be attributed to them. This aspect is crucial for complex systems with composite artificial intelligence systems that may have been developed by different entities.

Identification or registration, solutions which are regularly proposed for connected objects, are generally excluded from the debate on artificial intelligence because of the potential constraints for industrial and defence use, but this point is inherent to the effectiveness of a standard or certification based on ethics and accountability of the actors involved.

Registering artificial intelligence systems and establishing a mandatory signature for their actions could form part of the certification criteria. This would permit the creation of a register of certified artificial intelligence applications which could each be assigned a unique signature. This could be based on the model of the IP addresses allocated by ICANN's IANA for websites and code signing tools could be adapted to such applications. Using blockchain could amplify the register's transparency and strengthen its security, which would in turn enhance user confidence.

II.6 Indicators for Principle no. 5: under user control

> Principle 5 no. of the Charter "'Under user control': Preclude a prescriptive approach and ensure that users are informed actors and in control of their choices."

The principle of control by the different types of users goes hand in hand with explainability and the issue of interfaces, particularly the user interface (UI) and user experience (UX).

- Ensuring the system is under user control

By ensuring the system is under user control, the aim is to limit the impenetrability of the platform's functioning by enabling the user to understand and use all the available functionalities.

User control may be facilitated by:

- a helpline to provide access to relevant support
- an intuitive and ergonomic interface
- comprehensive and precise user guides
- short tutorials on how to use each functionality
- user training

As part of the certification process, for example at the end of the sandbox phase and depending on the complexity of the platform and its interface, the certification body could recommend measures to be taken to achieve the aim of ensuring the system is under user control with a view to obtaining certification. Tests involving "average" users could be carried out to verify the effectiveness of the support systems by measuring their ability to produce the desired result.

- AI opt-out for defendant

The Charter advocates clearly informing the defendant in the event of the use of artificial intelligence to render a judicial decision and providing for the right to object and request that the case be heard directly by a court within the meaning of Article 6 of the ECHR.

- Clear information for the defendant on the use of artificial intelligence in his/her case
  - Checking of the presence of an information banner mentioning the use of artificial intelligence that must be clicked before accessing the service
- Defendant’s right to opt out of the use of artificial intelligence
  - Checking of the presence of a mechanism for expressing consent on the last line of the information banner
  - Checking of a notification system for the defendant’s decision and for effective rerouting to conventional proceedings before a court within the meaning of Article 6 of the ECHR

- Training and certification of members of the judiciary

In the event of a state body deploying an artificial intelligence solution to support judicial decision-making with a view to testing it before rollout, it would also be useful to certify members of the judiciary in addition to the certification of the platform as such. Such certification training could focus on understanding the functionalities and the critical method to be adopted with regard to the proposals of artificial intelligence when drafting an assisted judicial decision.

III. Certification authorities and methods

The main certification methods, depending on which authority is to be competent, are as follows:

- Self-assessment by artificial intelligence publishers
- Assessment by CEPEJ or institute linked to the Council of Europe
- Assessment by approved bodies
- Continuous assessment by artificial intelligence plugin developed for or by the CEPEJ
- Mixed assessment capable of evolving

The choice of certification method and authority is partly driven by other aspects of certification. Criteria likely to guide the choice of the preferred type of certification authority include: whether certification is mandatory or optional, the desired degree of control, the economic profitability of the certification procedure or, conversely, its loss-making nature, technical complexity, the existence of bodies with a proven track record of certifying digital systems according to ethical criteria, and the feasibility of automation.

III.1 Self-assessment by artificial intelligence publishers

The main advantage of self-assessment by artificial intelligence publishers and/or subcontractors is that it enhances accountability. The self-assessment aspects discussed in this study are mainly inspired by the methods implemented to ensure GDPR compliance for personal data storage through impact assessments and the use of independent experts. Such an approach helps to raise awareness among publishers and to encourage continuous vigilance, but it does not seem sufficient in a field as sensitive as that of artificial intelligence in the judicial sphere, especially when it comes to connectionist artificial intelligence systems.

Relying on self-assessment as the sole method of verifying artificial intelligence systems seems particularly hard to reconcile with the demands of certification, particularly if the latter is mandatory.

III.2 Assessment by CEPEJ or institute linked to the Council of Europe

Having an assessment conducted by the CEPEJ or an institute linked to the Council of Europe that is dedicated to granting certification could be an option. This model could be based on the method of awarding the certificate of suitability to the monographs of the European Pharmacopoeia with the support of the Council of Europe’s European Directorate for the Quality of Medicines and Healthcare (EDQM) and the European Pharmacopoeia Commission set up for this purpose by the Council of Europe’s Public Health Committee.

The conditions for awarding certification of suitability to the monographs of the European Pharmacopoeia are laid down in Resolution AP-CSP (07) 1 of the Public Health Committee of the Council of Europe of 21 February 2007. This document sets out the scope, to whom the certificate is delivered and the procedure for granting the certificate (submission of the dossier, acknowledgement of receipt, designation of assessors, assessment procedures, notification of the decision, follow-up to the certification of suitability and reference documents).

The advantages of direct or near-direct certification by the CEPEJ include the procedural framework conducive to certification closely in keeping with the spirit of the Charter's principles, the certification's increased range of influence, closer management of the potential impact of lobbying by actors and companies involved the AI sector, and the direct selection of the individuals in charge of the assessments and the vetting of their impartiality, integrity and competence. Direct or near-direct certification may also have greater legitimacy justifying its mandatory nature.

Providing certification implies running costs, which are greater if the procedure is managed directly. By setting a certification application fee that covers the procedural operating costs, however, it would be possible to cover operating costs and even turn a profit that could, for example, be assigned to improving regulatory control of the use of artificial intelligence in the judicial sphere, to developing a plugin for monitoring artificial intelligence applications or to adapting the certification to other sectors in which artificial intelligence might violate human rights, particularly by exacerbating discrimination.

III.3 Assessment by approved bodies

Approved bodies or certificating bodies are organisations specifically authorised to independently ensure the application of the specifications and procedure required for certification granted by a public agency.

They have significant expertise in conducting certification. Having approved bodies set up a certification procedure may be quicker than a procedure being conducted by a public authority creating a specific entity for this purpose. In addition, these bodies achieve an economy of scale for certification, thereby limiting procedural costs and generating operating profits.

However, if optional certifications are entirely delegated to approved bodies, they may become less attractive for entities wishing to obtain certification of renown for their product or service.

III.4 Continuous assessment by artificial intelligence plugin developed for or by the CEPEJ

As mentioned above (in sections I.3, II.2 and II.5), an artificial intelligence plugin developed to verify the continuous compliance of systems using artificial intelligence in the judicial sphere could make it possible to automate compliance monitoring, or even part of the assessment itself. A plugin of this kind could provide support to artificial intelligence officers conducting compliance monitoring and possibly take on at least part of their work. If it were developed, such a tool would generate development and maintenance costs and a significant amount of time would be required to prove its effectiveness and to gain sufficient hindsight on the difficulties people had encountered during the certification procedure and monitoring process.

III.5 Mixed assessment capable of evolving

A mixed assessment capable of evolving could take advantage of the different benefits mentioned for each of the assessment methods and certifying bodies.

An initial self-assessment could therefore be carried out by the publisher applying for certification in the form of the impact assessments referred to above (see II) and the subsequent certification compliance monitoring could be partly carried out by an artificial intelligence officer designated by the publisher.

A pragmatic option might also be to delegate a pre-certification procedure or the assessment of technical aspects (mainly transparency and the security of information systems) to the abovementioned approved bodies in order, for example, to enable the CEPEJ or its partner organisation to focus its assessment on the aspects strictly related to human rights by design and the fight against discrimination in artificial intelligence solutions in the judicial sphere.

Deploying an artificial intelligence plugin as a complementary tool could also be useful to limit the constraints related to compliance monitoring following certification. However, this tool can only be envisaged in a second stage, as it should be designed on the basis of impact assessments and reports from artificial intelligence officers.

IV. Governance structure

IV.1 Governance models

The governance model of the European Directorate for the Quality of Medicines and Healthcare could serve as a source of inspiration in the event of the creation of a team responsible for the certification of artificial intelligence applications in the judicial sphere in the case, for example, of a partial agreement on this issue within the Council of Europe[55].

An organisation chart for a certification department for artificial intelligence in the judicial sphere could therefore include four sections:

- New Dossier Evaluation Section
- Certification Support Section
- Inspection Section
- Research, Development and Training Section

The New Dossier Evaluation Section would be responsible for processing the dossier for an initial certification procedure, forwarding it to the Certification Support Section where appropriate, designating assessors, taking decisions and sending notifications of the decision to award or refuse certification.

The Certification Support Section could be divided into two, with one unit specifically responsible for sandboxing and the other acting as a secretariat to support applicants drafting impact assessments and provide information during the certification procedure. The secretariat would also be the designated point of contact for any approved bodies and partner institutions.

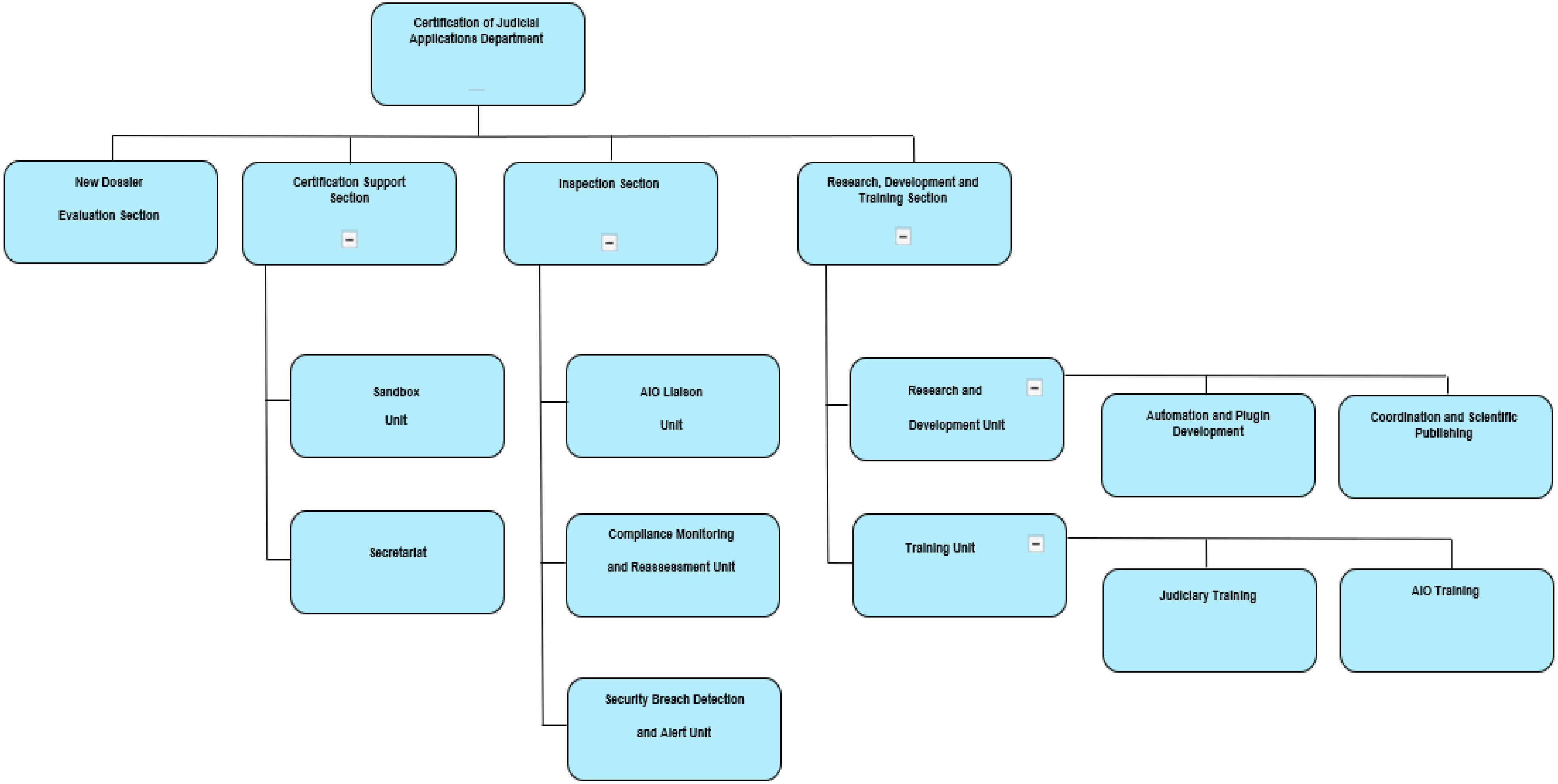
The Inspection Section could be split into three units. The first would be in charge of dealing with the artificial intelligence officers, the second responsible for compliance monitoring and reassessments and the third for detecting security breaches and issuing alerts. The Inspection Section would be responsible for continuous monitoring and could decide whether to suspend or withdraw certification.

The Research, Development and Training Section could be composed of two units with one responsible for research and development and the other for training. The Research and Development Unit could be subdivided into two parts, with one in charge of automation and the development of plugins to continuously monitor artificial intelligence systems and the other responsible for coordination and scientific publishing in order to monitor the latest developments, particularly in techniques for explaining artificial intelligence. It would also be in charge of conducting studies and looking at how to promote the certification beyond the judicial sphere. The Training Unit could be subdivided into two parts, with the first dedicated to the training and, if necessary, certification of members of the judiciary and the second to the training and, if necessary, certification of artificial intelligence officers.

[55] https://cs.coe.int/_layouts/15/orgchart/OrgChartCust_A.aspx?key=715&lcid=1033

IV.2 Possible organisation chart for a certification department for judicial artificial intelligence

The organisation chart below is adapted from that of the Certification of Substances Department (DECP) of the European Directorate for the Quality of Medicines and Healthcare (EDQM).



Certification of Judicial Applications Department
New Dossier Evaluation Section
Certification Support Section
Inspection Section
Research, Development and Training Section
Sandbox Unit
Secretariat
AIO Liaison Unit
Compliance Monitoring and Reassessment Unit
Security Breach Detection and Alert Unit
Research and Development Unit
Automation and Plugin Development
Coordination and Scientific Publishing
Training Unit
Judiciary Training
AIO Training

V. Identification and assessment of risks and opportunities entailed in certification by the CEPEJ

V.1 Competition risk and opportunities for cooperation with third-party certification projects

With the Charter and the certification project, the CEPEJ and the Council of Europe are leading a unique endeavour in the specific field of ensuring that artificial intelligence applications in the judicial sphere uphold human rights. In parallel, the Ad hoc Committee on Artificial Intelligence of the Council of Europe (CAHAI) examines the feasibility and potential elements of a legal framework for the development, design and application of artificial intelligence, based on Council of Europe's standards on human rights, democracy and the rule of law.

In the area of the ethics of artificial intelligence, other institutions, especially UNESCO and the OECD (see I.9), some Council of Europe member States such as Malta and Denmark (see I.7) and certification bodies such as CEN-CENELEC and IEEE (see I.8) have established principles that are in overall alignment with those of the Charter and they are considering certifying artificial intelligence applications.

The CEN-CENELEC and IEE could be partners of the future certification scheme, for example, in defining technical security standards adapted to the specific challenges of artificial intelligence in the judicial sphere.

Institutional cooperation could be considered with other intergovernmental and international institutions, such as UNESCO and the OECD, to examine how to propagate a human rights by design approach for artificial intelligence beyond the judicial sphere.

A consultation process on the strategies and first steps towards artificial intelligence certification taken by Council of Europe member States like Malta and Denmark could help to streamline proposals for certification methods.

V.2 Obsolescence risk

The certification's obsolescence risk seems limited insofar as it is designed to consider technological progress in general as well as the specific development of each certified artificial intelligence system. The obsolescence risk is also limited because the typologies used to define the indicators (symbolic and connective artificial intelligence) are based on their structure and are therefore fixed.

However, the obsolescence risk cannot be ruled out, for example, in the event of radically new technologies emerging or of notable advances in fields such as quantum computing, which are likely to considerably increase the power and calculation methods for quantum machine learning (QML)[56] and drastically modify standards in the field of information system security[57].

V.3 Risk of non-alignment with expectations of actors traditionally involved in technical standardisation

Technical standardisation like that of ISO operates largely on a model of consensus building between companies at successive levels: through national (e.g. AFNOR, DIN, etc.), regional (e.g. CEN) and then international (ISO) organisations.

The members of standards committees in the digital sector are either managers of start-up companies or small consulting firms or engineers seconded by large companies or national subsidiaries of multinational groups. The latter can pose a particular challenge to standard setting practices because the same multinational group may be represented in several national committees (multi-

[56] See for example: Vedran Dunjko, Hans Briegel, "Machine learning & artificial intelligence in the quantum domain: a review of recent progress", Rep Prog Phys, 2018, 81(7):074001.

[57] See for example: Dan Boneh, Mark Zhandry*, "Secure signatures and chosen ciphertext security in a quantum computing world"*, *Annual Cryptology Conference*, pp. 361-379, 2013; Marc Kaplan, Gaëtan Leurent, Anthony Leverrier, María Naya-Plasencia, "Breaking Symmetric Cryptosystems Using Quantum Period Finding", In: Robshaw M., Katz J. (eds) Advances in Cryptology –CRYPTO 2016. Lecture Notes in Computer Science, vol 9815. Springer, Berlin, Heidelberg.

representativity). Standards committees include very few legal professionals and very few academics. Technical standardisation is generally presented as a bottom-up process reflecting the priorities companies have identified in the field. The work of standards committees generally focuses on economically promising areas and interoperability issues. In the field of artificial intelligence, the relative economic insignificance of the judicial sphere (see introduction) means it is not considered in the discussions as a benchmark case. Certification resulting from the Charter therefore runs no risk of coming into conflict with the standardisation projects already under way.

National organisations also generally include ministry representatives who regularly chair the committee and sometimes steer certain standardisation work towards the strategic issues of their ministry. For example, in France, the AFNOR national committee on artificial intelligence is chaired by the Ministry of Defence and part of the work is therefore oriented towards autonomous weapons, governance and digital territory strategy.

At the European level, CEN-CENELEC launched a focus group on artificial intelligence in April 2019 to meet the standardisation needs identified by the European Commission[58]. The focus group explores different sectors such as smart manufacturing, robotics, autonomous transportation, virtual reality, health care, but applications in the judicial sphere are only mentioned in connection with the field of artificial intelligence-assisted decision-making. Strengthening relations with CEN-CENELEC and the presence, for example, of a CEPEJ representative as an observer, could make it possible to orient certain standardisation work towards the issues linked to certification on the basis of the Charter.

It should also be noted that the planned certification on the basis of the Charter is only partially technical and that the technical aspect focuses on ethics by design and human rights by design, in which may not be entirely compatible with a consensus-building process dominated by commercial companies. While cooperation on standardisation and certification work in this field could be established and serve a purpose, it would seem that such cooperation would have to focus on the technical aspect. The fact is that the points of divergence between the protection of fundamental rights and the drive for commercial profitability seem to be critical impediments to cooperation on the standardisation of ethics by design.

V.4 Risk of being classified as barrier to entry (WTO) and opportunities for ethical reorientation of practices

The introduction of technical standards or certifications that are difficult to attain or mandatory (see above I.1) may act as a brake on competition by making it difficult for new companies to enter the market, particularly when the standard or certification is mandatory or semi-mandatory. The risk of the World Trade Organisation (WTO) classifying such technical standards and certifications as a non-tariff barrier to market entry should also be considered. In theory, this has to some extent been the case for the Ecolabel scheme, although the certification's voluntary nature means a complaint against Ecolabels before the WTO on this basis would be unlikely to succeed[59].

Beyond its voluntary nature, the certification's origin plays an important role in the analysis of the risk of it being classified as a barrier to entry, which is less likely to be an issue for certification issued by an international or intergovernmental institution.

Furthermore, objective, neutral certification with a legitimate societal objective which seeks to ensure the enforcement of human rights could help avoid socially irresponsible corporate strategies and behaviour and promote ethical solutions designed with fundamental rights and freedoms in mind.
Therefore, the risk of artificial intelligence certification in the judicial sphere granted by the Council of Europe being classified as a barrier to market entry seems particularly low.

[58] See press release: https://www.cencenelec.eu/news/articles/Pages/AR-2019-001.aspx

[59] Kirstian Bartenstein and Sophie Lavallée, "L'écolabel est-il un outil du protectionnisme 'vert'?", Les Cahiers de droit, 2003, 44 (3), 361–393; Sophie Lavallée and Kristin Barentsein, "La régulation et l'harmonisation internationale des programmes d'écolabels sur les produits et les services", Revue internationale de droit économique 2004/1 (t. XVIII, 1), pages 47-77.

V.5 Opportunities for the theoretical-practical development and promotion of the human rights by design approach

The notion of Human rights by design laid down in the first principle of the Charter, of which it is an integral component, is meaningful and seems much more feasible than the widespread notion of ethics by design. Human rights by design differs from ethics by design in that it is based on provisions, articles to which methodological positivism can give real effectiveness.

The need to take into account human rights by design results from socio-technological factors. The human rights by design method should be considered as an extension to all the fundamental rights and freedoms of the privacy by design approach which shapes regulations governing the processing of personal data. The "by design" approach is made necessary by the place occupied in our societies by digital devices and algorithms which apply programming rules to the letter, unlike judges, for example, who at least implicitly take considerations of fairness into account. The words of Jean Bodin, "law without equity is like a body without a soul"[60], take on new meaning today in the era of machines enforcing rules with the precision which is their own. The method of human rights by design aims to ensure an application has built-in safeguards against the inflexibility of machine rules producing effects contrary to fundamental rights and freedoms. The method adopted for enforcing human rights by design could be inspired by the ways in which case-law contributions are integrated into codified legal texts. This approach corresponds to the integration of the *aequitas cerebrina*, cerebral or unwritten equity, into the *aequitas scripta*, written equity, that which is formalised in the law, in accordance with the *summa divisio*, the basic distinction, of the glossators[61]. Equity is traditionally used as a corrective mechanism, whereas the human rights by design approach aims to remove the need for subsequent correction as much as possible.

The human rights by design approach consists in anticipating dysfunctions that may appear in practice, correcting certain biases, rebalancing and formalising exceptions to the rule in a detailed manner. Impact assessments focusing on the protection of fundamental rights and freedoms are part of this human rights by design approach, in that they aim to anticipate the risks of implementation.

If successfully applied to artificial intelligence in the sphere of justice, the human rights by design approach could be applied more widely to artificial intelligence systems used in other fields and could also potentially be tested in legislative processes, after enactment of legislation, to limit the need for the review of compatibility of domestic legislation with international conventions and treaties before courts of law.

VI. Certification and responsibilities

At the CEPEJ-GT-QUAL meeting of 18 June 2020, it was considered that the accountability issues related to awarding certification were not particularly specific to artificial intelligence certification in the judicial sphere.

The generic liability issues can be broken down as follows:

- Placing responsibility on or removing responsibility from the creators of artificial intelligence in the judicial sphere
- Institutional and moral responsibility of the CEPEJ and the Council of Europe in the event of certification of platforms with shortcomings
- Responsibility for refusal of certification

Some liability issues of a slightly less generic nature can be highlighted such as the responsibility for a malfunction of the plugin deployed by the CEPEJ to certify artificial intelligence applications on a continuous basis or the limitation of the CEPEJ's responsibility through the use of a third-party certifier for strictly technical aspects, in particular those related to cybersecurity.

[60] Jean Bodin, Les six livres de la République de Jean Bodin (Six Books of the Commonwealth) (1530-1596).

[61] See for example: "L'équité ou les équités (Journées juridiques franco-libanaises, Paris, 3-4 October 2002)". In: Revue internationale de droit comparé. Vol. 55 N°1, January-March 2003. pp. 214-229.

VI.1 Implications of and responsibility for a malfunction of the plugin deployed by the CEPEJ

The option of deploying a plugin to continuously monitor artificial intelligence applications on legal and judicial platforms would enable increased compliance with the Charter post-certification and could have a potential impact beyond this sphere. Even if such a plugin were based, for example, on a symbolic artificial intelligence system and transcribed the criteria resulting from impact assessments and reports by artificial intelligence officers as a decision tree, malfunctions or security vulnerabilities could occur and have a significant impact on the user's perception, particularly for legal professionals. If the plugin malfunctioned and the monitored platform was no longer in compliance even though it appeared to be and was certified as such at the time, then a decision taken by judges using this tool would be undermined, potentially significantly reducing confidence in the system and the certification process and creating institutional tensions.

VI.2 Limiting responsibility through the use of a third-party certifier

The use of a third-party certifier (an approved body) to certify technical aspects, particularly those related to computer security, prior to full certification might be appropriate owing to the rapid pace of development of cutting-edge technology and in standards in this field. Such an approach would make it possible to ensure the quality of the certification and to relieve any certification body set up within the Council of Europe of ancillary tasks. It would also have the effect of limiting the direct responsibility of the CEPEJ and the Council of Europe in the event of inadequate certification in particularly technically complex areas that are not strictly within the scope of their missions and the added value of human rights certification by design.

VII. Points of convergence with the future European Union regulations on artificial intelligence

The European Union, and the European Commission in particular, are both very active in the field of artificial intelligence, whether it is by trying to make European companies competitive in the sector, by financing them (for example through its Horizon projects) or by setting the regulatory guidelines for artificial intelligence systems that uphold European values.

The White Paper on "Artificial Intelligence — a European approach based on excellence and trust"[62], published in 2020, is one of the first legislative steps towards imposing an ethical framework on artificial intelligence and the most notable. It follows on from a European Commission Communication entitled "Building Trust and Confidence in Human Driven Artificial Intelligence"[63] published in 2019 and is linked to the work of the High-Level Expert Group on Artificial Intelligence (AI HLEG)[64].

The European Commission's work proposes to regulate and impose a mandatory framework for "high-risk" artificial intelligence applications, while other artificial intelligence systems can be used with a certain degree of freedom and obtain optional certification by complying with various rules. An artificial intelligence application must meet two cumulative criteria to be considered high-risk: its use in a sector of probable risk and the consequences of such use. In the first criterion, the use of artificial intelligence in the "judiciary" is considered as a high-risk part of the public sector[65]. In the second criterion, the White Paper specifically mentions "uses of AI applications that produce legal or similarly significant effects for the rights of an individual or a company", which is the case, at least indirectly, for the large majority of artificial intelligence applications in the judicial sphere given their very purpose.

The European Commission therefore advocates an objective, prior conformity assessment to ensure compliance with mandatory requirements for high-risk artificial intelligence applications[66]. This approach

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

[63] Communication from the Commission to the European parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, "Building trust in human-centric artificial intelligence", Brussels, COM(2019)168 final, 8 April 2019.

[64] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[65] White Paper on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, p.17, footnote 50.

[66] Ibid. p. 23.

seems to be perfectly aligned with the CEPEJ's certification project which focuses on human rights by design and the definition of objective certification indicators.

The modalities of conformity assessment described by the European Commission refer to existing mechanisms such as CE marking[67], without excluding new, proportionate, non-discriminatory, transparent and objective mechanisms[68]. The main disadvantage of CE marking is that it is designed for finished products and not for systems that are likely to evolve such as connectionist artificial intelligence solutions. The CE marking mechanism, unless it is specifically adapted, lends itself rather poorly to impact assessments, sandboxes, the mandatory involvement of artificial intelligence officers and the use of a real-time certification monitoring plugin. Furthermore, opting for CE marking would have the effect of giving approved bodies discretion in points that go beyond the strictly technical and affect fundamental rights (see I.5 above).

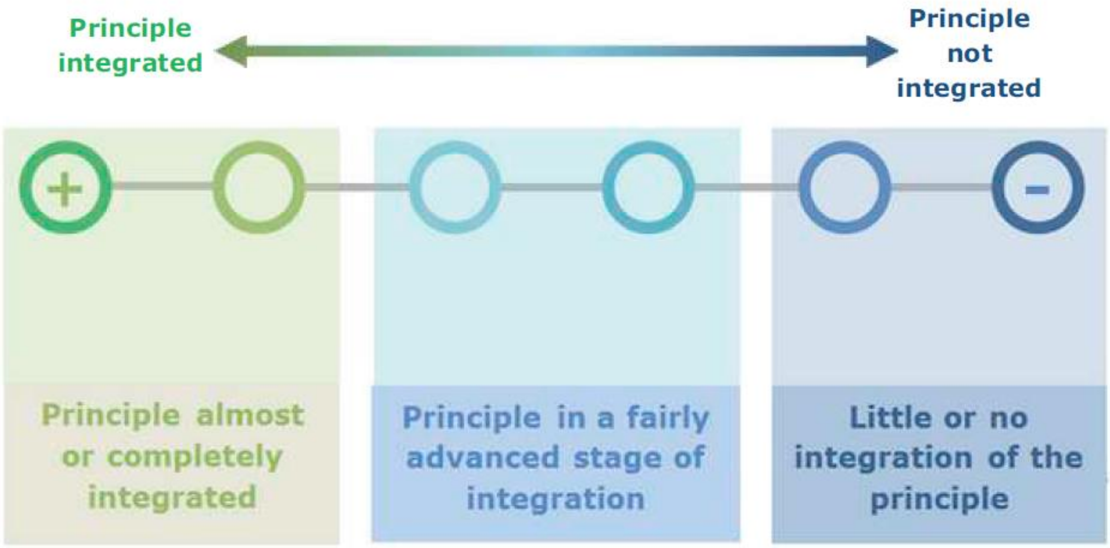
Overall, the European Commission's approach could converge with a mandatory certification project, which would be more appropriate than an optional certification project in the judicial sphere[69].

In addition, it should be noted that the European Commission is also considering support for SMEs to relieve the financial burden of applying for certification and seeking advice on compliance[70]. Such support could be particularly appropriate for the implementation of sandboxes (see I.11).

VIII. Schedule and roadmap

VIII.1 Ramification with the "Checklist of Charter principles in your processing"

Appendix IV to the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems includes a "Checklist for integrating the Charter's principles into your processing method". The checklist is based on self-evaluation of the degree of integration of each principle in the artificial intelligence system.


Principle integrated
Principle not integrated
+
-
Principle almost or completely integrated
Principle in a fairly advanced stage of integration
Little or no integration of the principle

(CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems, Appendix IV, "Checklist for integrating the Charter's principles into your processing method", p 76)

[67] Ibid. p.23 "The conformity assessments for high-risk AI applications should be part of the conformity assessment mechanisms that already exist for a large number of products being placed on the EU's internal market".

[68] Ibid. p.23 "Where no such existing mechanisms can be relied on, similar mechanisms may need to be established, drawing on best practice and possible input of stakeholders and European standards organisations. Any such new mechanism should be proportionate and non-discriminatory and use transparent and objective criteria in compliance with international obligations".

[69] *Idem.*

[70] *Idem.*

It is specified in Appendix IV that this checklist by no means equates to the issue of a label or certification.

This checklist could be integrated at the stage of the impact assessment by the actors involved, upstream of the labelling process.

VIII.2 Definition of initial deployment needs and milestones for the CEPEJ

To allow the deployment of the certification project, the creation of a dedicated team within the secretariat of the CEPEJ could be envisaged.

Initially, during a preliminary phase to the deployment of a certification, the necessary human resources may be limited and focus on two objectives : on the one hand the deployment and institutional follow-up and on the other hand the training and technical aspects.

For the deployment and the institutional follow-up, a coordinator assisted by a restricted team of one to two persons, could be in charge of the progress of the certification project, of the internal liaison within the CEPEJ and the Council of Europe and of the inter-institutional and public communication relating to the certification project. The coordinator could be in charge of the follow-up of a call for expression of interest and the consultation of governmental and non-governmental bodies, actors in the sector (legal professionals, platform publishers and artificial intelligence developers) and the public.

The coordinator's profile should be specialized in international and institutional relations, ideally with experience in a certification body, an international organization focused on digital or experience in digital law (IT lawyer, lawyer in a digital company ...). Assistants, juniors or interns, could have training in communication or public relations, ideally with prior experience in intergovernmental organizations and/or digital communication agencies. Certain aspects of communication could be outsourced or could be carried out using pre-existing CEPEJ resources.

For the training and technical aspects, a team could be composed of two complementary profiles: on the one hand, a magistrate or teacher-researcher in digital law, in charge of developing a training system for magistrates and delegates in artificial intelligence, and on the other hand, a standardization engineer specialized in information systems, in charge of the technical follow-up of the certification project and the technical aspects of the training. Academics and technical experts from certification bodies could be called upon from time to time to contribute to the work of the training and technical team.

In a second phase, once the deployment has been completed, additional human resources will be required for each of the sections: evaluation of new files, support for certification, inspection. The team already created dedicated to training could be maintained and possibly restructured to respond more specifically to the needs identified. These additional recruitments could be progressive, depending on the number of applications for certification, which would allow an adjustment with the application fees charged to applicants for certification.

Conclusion

By drawing up the Charter and seeking a framework or certification procedure to promote the enforcement of fundamental rights and freedoms in artificial intelligence applications, the Council of Europe is aligning its mission to raise awareness and ensure the protection of human rights with technological progress by adapting to this change and shaping its development to respect societal needs.

The use of artificial intelligence in the legal and judicial sphere represents an important societal challenge, particularly in the field of algorithmic justice. While technological advances may bring improvements to the judicial system, facilitate the work of legal professionals and improve access to justice and information for defendants/litigants, increased vigilance is necessary in this sector which the European Commission describes as a high-risk sector.

The creation of objective, neutral certification aimed at enforcing human rights by design would foster the emergence of an ecosystem of artificial intelligence applications designed and deployed in a manner that upholds fundamental rights and freedoms. In a high-risk sector, such as the judicial sphere, the mandatory nature of certification appears to garner consensus, but its implementation should not impede innovation.

The CEPEJ's certification initiative is both unprecedented from a sectoral point of view and convergent with the work on ethics and artificial intelligence under way in other bodies and institutions such as UNESCO and the OECD and responds to a need already expressed by several Council of Europe member States. It could also build on certain technical aspects of the standardisation groundwork being carried out, for example, by CEN-CENELEC and IEE.

The certification of artificial intelligence systems in the judicial sphere would also make it possible to support private and public projects and to establish standards that reach beyond Europe, justifying, for example, the development of international mechanisms for the recognition and enforcement of foreign decisions[71] or arbitral awards[72] made by or with the assistance of artificial intelligence.

A successful experience in certification of artificial intelligence in the judicial sphere, where ethical considerations and fundamental rights and freedoms are essential, could serve as a useful source of inspiration for artificial intelligence certification in other fields.

[71] For example, Regulation (EU) No 1215/2012 of the European Parliament and of the Council of 12 December 2012 on jurisdiction and the recognition and enforcement of judgments in civil and commercial matters.

[72] Convention on the Recognition and Enforcement of Foreign Arbitral Awards, New York, 1958.

Appendixes

Summary table of indicators and certification criteria

Objective	Criteria	Assessment method	Target	AI category
Proportionate processing of personal data	Anonymisation of the parties and participants (physical individuals) and their counsels	Consultation of data sets	Unprocessed data	All
	Absence of evaluation and classification of physical individuals or legal entities on the basis of judicial decisions	Checking of interface	Interface	Connectionist
		Checking of database		
Limit forum shopping	Anonymisation of the judge and the court's location in decisions used for predictive justice	Consultation of data sets	Unprocessed data	Connectionist
Clear purposes for processing	Hermetic separation of artificial intelligence services	Checking of databases and data sources used by each system	Databases	All
Fair trial	Information indicating to the judge and the defendant, if relevant that a report generated by an artificial intelligence system is not explainable (See also below : Defendant's/litigant's right to opt out of the use of artificial intelligence)	Checking of AI category and checking of existence and clarity of information (See also below : Checking of a notification system for the defendant's/litigant's decision and for effective redirection to conventional proceedings before a court within the meaning of Article 6 of the ECHR)	Learning model and interface	Connectionist
Judges' independence in their decision-making process	Safeguard against the profiling of judges	A/B testing checking of search results	Search engine and processed data	All
	Match between the criterion displayed and the actual pattern of classification of search results	Auditing of search results	Search engine	All
	Transparency of weighting of criteria for multicriteria searches	Checking of the existence of explanatory information and auditing of search results	Interface and search engine	All
		Verification by auditing of search results		
		Checking of the existence of	Interface and search engine	All

Summary table of indicators and certification criteria

	Transparency of criteria used for searches by "relevance"	explanatory information and auditing of search results		
		Verification by auditing of search results		
Ethics and Human rights by design	No human rights violation	Report presenting decision trees and explaining how fundamental rights and freedoms are taken into account	Report	Symbolic
	No human rights violation	Report presenting training data and methods and explaining how fundamental rights and freedoms are taken into account	Report	Connectionist
Avoiding discrimination based on sensitive data	Elimination of the tags that could be linked to parties' sensitive data (home address, income, family situation, registered capital)	Checking by consultation of data sets	Unprocessed data	Not under public authority control
		A/B testing using information and tags that could be linked to sensitive data by changing, where applicable, one of the following parameters during each test: name, home address, income, family situation, registered capital, relevant specific contextual information, etc.	Search engine	
		Form enabling users to submit detailed requests for the removal of tags that could be linked to sensitive data, with copy to the artificial intelligence officer, where applicable, and to the control authority (labelling entity	Interface	
	Mechanisms guaranteeing that the processing is necessary and proportionate	Impact assessment by the processing body explaining the safeguards implemented	Report	Under public authority control

Summary table of indicators and certification criteria				
		Designation of an artificial intelligence officer	Letter of designation	
Data security and quality	Prior certification to ISO/IEC 27001 Standard	Checking of certification to ISO/IEC 27001 Standard	Certification by an accredited certification body	All
Transparency	Open source or transmission of source code on a confidential basis	Checking of entire source code	Source code	All
Ensuring the system is under user control	Helpline providing relevant support; intuitive and ergonomic interface; comprehensive and precise user guides; short tutorials on how to use each functionality; user training	Checking of presence on interface	Interface and functionality test	All
User control and fair trial	Clear information for the defendant/litigant on the use of artificial intelligence in his/her case	Checking of the presence of an information banner mentioning the use of artificial intelligence that must be clicked before accessing the service	Interface	All
	Defendant’s/litigant's right to opt out of the use of artificial intelligence	Checking of the presence of a mechanism for expressing consent on the last line of the information banner	Interface	All
		Checking of a notification system for the defendant’s/litigant's decision and for effective redirection to conventional proceedings before a court within the meaning of Article 6 of the ECHR	Interface and functionality test	

Additional tools

Description	Objective
Interpretable artificial intelligence plugin to monitor artificial intelligence applications in the judicial field	Continuous monitoring of connectionist AI systems and support for AI officers
Blockchain ledger to register and provide signature for certified artificial intelligence systems	Creating records to enable identification of AI system
Blockchain ledger for judicial decisions	Data integrity and compliance with anonymisation requirements
Sandbox	Supporting applicants during certification process
Practical training for members of the judiciary on artificial intelligence in the judicial sphere	Raising awareness about challenges related to the Charter and to certification, acquiring technical skills and sharing best practices