8 décembre 2020

CEPEJ(2020)15Rev

COMMISSION EUROPEENNE POUR L'EFFICACITE DE LA JUSTICE

(CEPEJ)

Mise en place éventuelle d'un mécanisme de certification des outils et services d'intelligence artificielle dans le domaine juridique et judiciaire :

Etude de faisabilité

En décembre 2018, la Commission européenne pour l'efficacité de la justice (CEPEJ) a adopté la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement. La Charte de la CEPEJ représente un premier pas dans les efforts de la CEPEJ pour promouvoir l'utilisation responsable de l'intelligence artificielle (IA) dans les systèmes judiciaires européens, conformément aux valeurs du Conseil de l'Europe. Conscient de la nécessité de soutenir la mise en œuvre de la Charte, le Groupe de travail de la CEPEJ sur la qualité de la justice (CEPEJ-GT-QUAL) a mené une réflexion sur la mise en place éventuelle d'un mécanisme de certification des solutions d'IA en conformité avec les principes de la Charte.

Cette étude de faisabilité a été élaborée, sous la supervision du CEPEJ-GT-QUAL, par M. Matthieu Quiniou (France), expert scientifique, Avocat à la Cour d'Appel de Paris et Enseignant-chercheur (Université Paris 8).

Telle qu'adoptée lors de la 34ème réunion plénière de la CEPEJ,
8 décembre 2020

Table des matières

Etude de faisabilité sur la mise en place éventuelle d'un mécanisme de certification des outils et services d'intelligence artificielle

Introduction

La justice n'est pas actuellement le secteur de prédilection des entreprises innovantes en matière d'intelligence artificielle comme l'illustre, par exemple, le classement Forbes 2019 des cinquante entreprises d'intelligence artificielle les plus prometteuses, dans lequel ne figurent aucune entreprise rattachable aux domaines juridique ou judiciaire[1]. Cette situation peut s'expliquer par la spécificité, la complexité, les barrières réglementaires à l'accès du marché et également la relativement faible importance économique du marché du droit. En effet, le marché du droit à l'échelle mondiale représente à peine plus de 1% du PIB annuel mondial, soit environ 1 trillion[2] sur les 80 trillions d'euros du produit mondial brut estimés par le Fond Monétaire International[3]. Une étude française consacrée au poids économique du marché du droit indiquait à titre illustratif que le marché du droit est équivalent à celui du transport aérien, de la publicité ou encore des boissons[4]. Pour autant, l'intelligence artificielle dans le domaine juridique est un enjeu sociétal de premier ordre dès lors qu'il s'agit d'anticipation de délits ou d'automatisation des décisions de justice.

Actuellement plusieurs entreprises innovantes du secteur juridique, souvent appelées Legal Tech et certains Etats proposent des services utilisant de l'intelligence artificielle pour améliorer et personnaliser les résultats de recherche sur des bases de données juridiques et pour faciliter la prise de décision notamment en matière de provision dans un contexte contentieux. Les dispositifs les plus ambitieux proposent même de calculer les risques de récidives, même si certains de ces outils déjà expérimentés ont parfois été écartés pour des raisons d'éthiques mais également d'inefficacité, comme dans le cas du système COMPAS aux Etats-Unis[5].

Cette étude s'appuie principalement sur la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ (ci-après « Charte »), dont elle est le prolongement direct, mais également sur les travaux suivants :

- Etude sur les dimensions des droits humains dans les techniques de traitement automatisé des données (en particulier les algorithmes) et éventuelles implications réglementaires, Etude du Conseil de l'Europe, DGI(2017)12 préparée par le comité d'experts sur les intermédiaires d'internet (MSI-NET) ; et
- Livre Blanc, Intelligence artificielle Une approche européenne axée sur l'excellence et la confiance, COM(2020) 65 final.

Cette Etude a également pris note des travaux en cours du Comité ad hoc sur l'Intelligence artificielle (CAHAI), récemment créé par le Comité des Ministres du Conseil de l'Europe.

L'étude approfondie sur l'utilisation de l'IA dans les systèmes judiciaires, notamment les applications d'IA assurant le traitement des décisions et des données judiciaires, reproduite en Annexe n° 1 de la Charte de la CEPEJ[6], précise à titre illustratif les grandes familles d'intelligence artificielle en matière de justice :

- Moteur de recherche de jurisprudence avancé
- Résolution de litiges en ligne,
- Aide à la rédaction d'actes
- Analyse (prédictif, barèmes),
- Catégorisation des contrats selon différents critères et détection de clauses contractuelles divergentes ou incompatibles,
- « Chatbots » de renseignement du justiciable ou de support de celui-ci dans sa démarche litigieuse.

[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 et 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] Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ, Annexe 1, page 17.

Deux catégories supplémentaires à fort enjeu éthique peuvent également être envisagées, celle de la justice algorithmique, qui pourrait être rapprochée de la catégorie susvisée de « Résolution de litiges en ligne » et la catégorie d'outils d'aide à la décision pour le juge augmenté, pouvant être rapprochée de la catégorie « Analyse », encore à un stade expérimental et principalement axée sur l'aide à la détermination des dommages-intérêts et sanctions.

Du point de vue de la certification, ces solutions technologiques sont suffisamment différentes dans leur objet et dans l'usage qui est fait de l'intelligence artificielle qu'il pourrait être envisagé soit de ne certifier qu'une catégorie de ces usages soit de prévoir plusieurs sous-catégories de certifications avec des procédures et critères différenciés.

Lors de la réunion de la CEPEJ-GT-QUAL du 18 juin 2020, les cas d'usages assimilables à de la justice prédictive ont été mis en avant comme ceux susceptibles d'avoir les incidences les plus importantes en matière de droits et libertés fondamentaux et donc à privilégier dans le cadre de la présente étude.

A cette diversité d'avancement technologique, de fonctionnalités et d'usages de ces intelligences artificielles en matière judiciaire, il convient d'ajouter la prise en compte des domaines du droit et des particularités juridictionnelles par champ du droit. Il convient tout d'abord de préciser que certains domaines du droit sont plus sensibles que d'autres, en raison des conséquences des décisions de justice sur les libertés. La criminologie et le droit pénal sont ainsi généralement présentés comme des domaines du droit pour lesquels l'utilisation de l'intelligence artificielle doit être envisagée avec la plus grande prudence. En outre, pour prendre le cas particulier de la France, les décisions sont rendues selon la matière par des tribunaux composés exclusivement de magistrats professionnels, par des juges référendaires non-professionnels, par des représentants syndicaux ou encore par des jurés populaires. D'autres systèmes judiciaires, souvent anglo-saxons, mettent un point d'honneur à permettre la transcription d'opinions divergentes de juges. Cette diversité des juges est source de créativité et permet des ponts entre la société et le droit. De nombreux rapports mettent en évidence le risque d'une sclérose de la jurisprudence du fait d'une restitution stricte par l'intelligence artificielle des positions figées par les juridictions suprêmes (dont le rôle est d'ailleurs celui de l'harmonisation d'un droit en mouvement).

La principale typologie des intelligences artificielles distingue entre celles dites symboliques (ou cognitivistes) et celles dites connexionnistes (ou numériques). Les intelligences artificielles dites symboliques s'appuient sur des règles fixées par leur concepteur pour permettre à la machine de prendre des décisions à partir d'un modèle pré-établi. Utilisée depuis de nombreuses années dans les systèmes experts, cette intelligence artificielle est principalement utilisée dans le domaine juridique pour les agents conversationnels (*chatbot*) et permet également de réaliser de la programmation par contrainte, par exemple avec des arborescences permettant de générer automatiquement des contrats complexes à partir de questions simples. Les intelligences artificielles dites connexionnistes sont probabilistes et fonctionnent par induction et itération pour déduire des données des caractéristiques susceptibles de devenir des indicateurs et d'interpréter des jeux de données. Pour entraîner les intelligences artificielles dites connexionnistes, il est nécessaire de disposer à fois de bases de données importantes (*big data*) et d'une puissance de calcul suffisante pour traiter les données dans un délai limité. Les intelligences artificielles dites connexionnistes sont celles qui permettent par exemple de définir une stratégie contentieuse ou d'estimer une sanction à partir d'une jurisprudence importante.

Les droits et libertés fondamentaux relèvent d'une ontologie humaine, différente de l'approche fonctionnelle des machines, même de celles dites intelligentes. A la différence des objets inanimés ou des machines avec des comportements préprogrammés ou strictement délimités, l'intelligence artificielle se définit justement par le dépassement du modèle statistique proposé et son amélioration constante par itérations successives.

La certification d'une intelligence artificielle connexionniste peut être réalisée à plusieurs niveaux : au niveau du modèle d'apprentissage (supervisé ou non, validité scientifique du protocole et biais d'apprentissage), au niveau des données utilisées (validité de la donnée, sécurisation de la base de données et exclusion de certaines données sensibles) et au niveau des résultats (impact négatif sur les droits et libertés fondamentaux).

Les intelligences artificielles, les plus évoluées, intelligences artificielles connexionnistes à apprentissage non supervisé, adaptatives (non déterministes), traitent les données de manière contextualisée ce qui rend les discriminations portant atteinte aux droits et libertés fondamentaux plus difficiles à discerner pour des êtres humains non assistés par une méta-intelligence artificielle entraînée spécifiquement à cette tâche. Par ailleurs, le caractère constamment évolutif de l'intelligence artificielle nécessite un suivi de certification continu, possible par exemple en adaptant le modèle des robots d'indexation et en créant des plugins.

Outre ces considérations liées aux catégories d'intelligence artificielle et d'avancement technologique de celle-ci en matière d'apprentissage contextuel, il convient de distinguer les intelligences artificielles en matière de justice selon leurs fonctions et finalités.

Dans la réponse du CEN-CENELEC de juin 2020[7] au Livre Blanc COM(2020) 65 final l'importance de la prise en compte des applications spécifiques et des enjeux sectoriels a été mise en avant en matière de normalisation et de certification des intelligences artificielles[8]. Tout en notant l'absence de normes stabilisée en matière d'intelligence artificielle, le CEN-CENELEC propose dans ses recommandations de s'inspirer de l'Ecolabel européen, de la certification en matière de sécurité en cours de finalisation et des labélisations nationales du Danemark et de Malte[9].

Les questions qui seront abordées prioritairement dans la présente étude concerneront la typologie et les enjeux des certifications et labels (I) et les objectifs de cette certification par la CEPEJ (II). Cette étude présentera également les enjeux liés au déploiement de cette certification, en termes d'autorité de certification (III), de structure de gouvernance (IV), ainsi que les risques et opportunités d'une telle certification par la CEPEJ (V) et les enjeux de responsabilités liés au déploiement de la certification (VI). Cette étude sera également mise en perspective par rapport à la future réglementation de l'Union européenne en matière d'intelligence artificielle (VII) et un calendrier de déploiement et une feuille de route seront suggérés (VIII).

I. Etat des lieux, typologie et enjeux des normes, certifications et labels

La certification éthique de l'intelligence artificielle apparaît comme un enjeu important dans les domaines stratégiques dans lesquels les boites noires algorithmiques sont démocratiquement inadmissibles. Des entreprises européennes de secteurs stratégiques se positionnent pour une intelligence artificielle certifiée fiable et explicable, souvent en opposition aux grandes entreprises américaines du numérique, les GAFAM[10]. Le besoin de certification par les entreprises européennes est perceptible, comme, par exemple, dans la déclaration du Président-directeur général du Groupe Thales, Patrice Caine selon lequel : « L'Intelligence Artificielle sera bientôt au cœur du quotidien de chaque citoyen. Pour nos clients dans les secteurs de l'aéronautique, de l'espace, des transports terrestres, de la défense et de la sécurité, qui ont en charge des systèmes critiques pour la sécurité de nos sociétés, s'assurer du bon fonctionnement de l'IA, en décrypter les mécanismes et en certifier l'utilisation, revêt une importance cruciale. C'est pour cette raison que Thales s'engage pour une IA de confiance, explicable, certifiable et responsable »[11].

I.1 Incidences du choix entre certification obligatoire et facultative

De manière générale les normes techniques ont pour principal objectif de répondre à des besoins du marché, notamment pour la compatibilité et les certifications ont également un effet d'incitation des acteurs à l'adoption de certains comportements. Pour les acteurs, les normes et certifications peuvent conditionner la mise sur le marché, les modalités de présentation du produit ou service et également les critères à remplir lors de réponses à des appels d'offre publics et parfois privés.

Les certifications comme les normes techniques pour les produits ou services sont par principe d'utilisation volontaire. La réglementation peut néanmoins rendre certaines normes techniques obligatoires, comme par exemple les directives de l'Union européenne relatives au marquage CE de certains produits.

Les normes techniques et certifications ont notamment pour fonction de permettre au consommateur d'identifier simplement des qualités d'un produit ou service dans le cadre d'une comparaison sur le marché, ce qui favorise en principe la concurrence. Pour autant, lorsque ces normes techniques ou certifications sont difficiles à obtenir, cela peut freiner la concurrence en rendant difficile l'accès au marché par de nouveaux acteurs, particulièrement lorsque la norme ou certification est obligatoire ou semi-obligatoire.

---

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

[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 et Microsoft.

[11] « Thales illustre le rôle crucial de l'Intelligence Artificielle dans les moments décisifs », site de Thales, 17 janvier 2019 https://www.thalesgroup.com/fr/group/press-release/thales-illustre-le-role-crucial-lintelligence-artificielle-moments-decisifs

Il convient également de noter qu'une labellisation obligatoire ou quasi-obligatoire empêchant des services utilisant des boites noires d'intelligence artificielle non interprétables ou ne respectant pas des principes éthiques d'intelligence artificielle pourrait avoir des effets très significatifs sur la possibilité pour de grandes entreprises du numérique de proposer leurs services sur certains territoires, sauf à revoir leur approche. Une telle labellisation ou certification pourrait donner des critères objectifs de protection des utilisateurs de ce type de services et limiter les confrontations diplomatiques, comme en 2020 entre les Etats-Unis et l'entreprise chinoise TikTok[12].

I.2 Différences entre certification d'intelligence artificielle symbolique et connexionniste

L'intelligence artificielle symbolique s'appuie exclusivement sur des modèles implémentés par leur concepteur, ce type d'intelligence artificielle est déterministe. L'intelligence artificielle symbolique s'appuie sur des arbres de décisions entièrement interprétables et auditables. Un critère discriminatoire pourra être identifié aisément et modifié, la certification de ce type d'intelligence artificielle est relativement simple à mettre en œuvre et peut être réalisée à partir de critères d'audits des bases de données, de la formulation des critères et de l'arbre de décisions.
A l'inverse la normalisation et la certification de l'intelligence artificielle connexionniste, tout particulièrement celles fonctionnant avec de l'apprentissage non supervisé, est complexe et encore à l'état d'ébauche. L'enjeu principal actuel est de permettre de comprendre précisément les étapes de raisonnement de la machine sans la brider dans sa possibilité de réaliser des interprétations inédites et pertinentes. Certains auteurs proposent de distinguer deux méthodes pour rendre l'intelligence artificielle interprétable, soit rendre la machine moins complexe soit appliquer une méthode qui analyse le modèle après la phase d'entrainement[13].

Actuellement, il n'existe pas de système permettant directement l'interprétation d'un modèle d'intelligence artificielle connexionniste à apprentissage non supervisé. Les chercheurs distinguent généralement entre explicabilité (désignée régulièrement sous l'acronyme XAI pour intelligence artificielle explicable), c'est-à-dire la possibilité de comprendre les effets généraux des variables (le pourquoi) et interprétabilité, c'est-à-dire la possibilité de quantifier l'importance de chaque variable dans le résultat (le comment) d'un algorithme.

Des méthodes comme la méthode LIME (*Local Interpretable Model-agnostic Explanations*[14]) sont utilisées pour faciliter la compréhension des résultats issus des boites noires de réseaux de neurones[15]. Des évaluations de la qualité de l'interprétation ont déjà été proposées et certains auteurs mettent en avant notamment des problèmes éthiques liés au fait de « produire des explications davantage persuasives que transparentes »[16].

Des travaux comme ceux du projet *TuringBox*[17] porté par des chercheurs du Massachusetts Institute of Technology (MIT) peuvent également être intéressants pour faciliter le travail d'interprétation des intelligences artificielles dans le domaine juridique. Ce projet permet à des personnes sans compétences spécifiques en informatique d'examiner une intelligence artificielle en la stimulant et en analysant des métriques. Un des principaux intérêts de cet outil est son caractère participatif permettant de partager des analyses réalisées.

Concrètement dans le domaine judiciaire avec des outils d'intelligence artificielle d'aide à la décision pour les juges, si, par exemple, la machine propose une mesure de détention plutôt qu'une mesure de sûreté, le justiciable et la société sont en droit d'exiger une explication à ce choix s'il guide la formalisation d'une décision de justice. Cette nécessité s'inscrit dans le prolongement de l'article 5 « Droit à la liberté et à la sûreté » de la Convention européenne de sauvegarde des droits de l'homme et des libertés fondamentales (ci-après « CESDH ») qui implique l'obligation de motivation des décisions et l'interdiction de l'arbitraire[18]. La motivation

---

[12] https://www.reuters.com/article/us-usa-tiktok-china-pompeo-idUSKBN2480DF
[13] Voir Christoph Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, lulu.com, 2020, 318 pages.
[14] Peut être traduit par « Explications locales interprétables par modèle-agnostique ».
[15] Voir par exemple, 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.
[16] Voir par exemple : Christophe Denis et 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), Juillet 2019, Toulouse, France. pp.60-68.
[17] https://turingbox.mit.edu/ et https://arxiv.org/abs/1803.07233
[18] Cour européenne des Droits de l'Homme, Guide sur l'article 5 de la Convention européenne des droits de l'homme Droit à la liberté et à la sûreté, màj 31 décembre 2019, p.15.

des décisions de justice impose un certain niveau d'explicabilité. Dans le cas de figure susmentionné d'une intelligence artificielle proposant une mesure de détention, en fonction du type d'intelligence artificielle (symbolique ou connexionniste, avec ou sans apprentissage supervisé) utilisé et des outils mis à la disposition du juge pour l'interprétation de l'intelligence artificielle, celui-ci se trouvera dans une des trois configurations suivantes. Soit le juge sera dans l'incapacité totale d'expliquer la décision de l'intelligence artificielle. Soit le juge ne pourra identifier qu'un critère irrationnel, sans pertinence ou insusceptible de fonder une motivation juridique. Soit le juge sera en capacité d'expliciter clairement la décision prise par l'intelligence artificielle. Dans les deux premiers cas, une vigilance accrue est nécessaire pour éviter une pratique consistant à chercher à atteindre le même résultat que l'intelligence artificielle en termes de sanction à partir d'un autre raisonnement, celui-ci explicable.

I.3 Certification d'un dispositif évolutif

L'intelligence artificielle connexionniste est par nature non déterministe, évolutive. Seuls certains invariants comme les sources de données ou le modèle initial d'entraînement (pour celles ayant un entraînement supervisé) peuvent être certifiés durablement. Les itérations successives de l'intelligence artificielle connexionniste ont pour effet de la complexifier et en principe de la rendre plus pertinente, pour autant, cette complexification rend l'explicabilité des résultats plus complexes.

Pour ce type d'intelligences artificielles non déterministes, la certification se doit d'être continue pour être efficace. Il est ainsi possible d'envisager un contrôle humain continu pour les intelligences artificielles certifiées, avec la création par exemple d'une fonction de délégué à l'intelligence artificielle au sein des entreprises ou des administrations proposant ce type d'intelligence artificielle certifiée sur le modèle du délégué à la protection des données (ci-après « DPO ») ou d'envisager un contrôle automatisé par une intelligence artificielle symbolique, sous forme de plugin, vérifiant que l'intelligence artificielle connexionniste qu'elle contrôle remplit toujours les critères ayant justifié initialement sa certification.

I.4 Comparaison avec la certification des dispositifs de traitement des données personnelles

La certification de l'intelligence artificielle repose sur l'analyse de l'algorithme, du modèle d'entraînement et de la qualité des sources de données. Les certifications des dispositifs de traitement de données personnelles ont pour principaux objectifs de vérifier le respect des réglementations sur les données personnelles, dans l'Union européenne, le Règlement général sur la protection des données (ci-après « RGPD »)[19], ce qui se traduit principalement par une certification des modalités d'information et d'obtention du consentement des personnes dont les données sont collectées et traitées et, de manière résiduelle, une certification des précautions et modalités de sécurisation de ces données.

En matière de données personnelles depuis l'entrée en vigueur du RGPD, les autorités indépendantes de protection des données nationales (membres du G29), comme en France la CNIL, ont renoncé aux labels liés aux procédures de gouvernance tendant à assurer la protection des données, au label de service de coffre-fort numérique et aux labels de procédure d'audit, pour se concentrer uniquement sur un référentiel de certification des DPO et des organismes certificateurs[20]. En l'état actuel, ces certifications ne peuvent être des sources d'inspiration que dans l'hypothèse de la création de la fonction susmentionnée de déléguée à l'intelligence artificielle. Il convient de noter que des certifications supplémentaires devraient certainement apparaître dans différents Etats membres de l'Union européenne, l'article 42 du RGPD « Certification » encourage à des certifications et labels aux fins de démontrer que les opérations de traitement effectuées respectent le RGPD.

En matière de protection des données personnelles, il existe également des normes ISO, tout particulièrement la norme ISO/IEC 27701. Cette norme internationale présentée par l'ISO comme la première norme internationale sur le management de la protection de la vie privée[21] a été publiée en 2019. Cette norme correspond à l'état de l'art de la protection de la vie privée et regroupe notamment les exigences relatives à la création, la mise en œuvre et l'amélioration des systèmes de management de la protection de la vie privée (*Privacy Information Management System* - PIMS)[22] et en matière de sécurité. Cette norme peut être pensé comme un préalable et une composante d'une certification portant sur un système d'intelligence artificielle. Il convient de noter que cette norme ISO/IEC 27701 n'est pas une certification RGPD au sens de l'article 42 du

[19] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données).
[20] Par exemple en France pour la CNIL, AFNOR ou Bureau Veritas.
[21] https://www.iso.org/fr/news/ref2419.html
[22] https://www.iso.org/fr/standard/71670.html

RGPD et s'inscrit dans ce qui a pu être défini comme une zone grise de normalisation avant la formalisation de certifications issues des autorités réglementaires[23].

I.5 Mise en perspective du marquage CE pour des systèmes d'intelligence artificielle

Le marquage CE est encadré par le règlement (CE) no 765/2008 quant à sa définition, à son format et à ses principes généraux. Le marquage CE, n'est pas une certification mais un engagement visible d'un fabricant à respecter la législation européenne. Le guide bleu de la commission européenne indique ainsi : « en apposant le marquage CE, le fabricant déclare sous sa seule responsabilité que le produit est conforme à l'ensemble des exigences législatives applicables de l'Union, et que les procédures d'évaluation de la conformité appropriées ont été appliquées avec succès »[24]. Le marquage CE n'est obligatoire que pour certains produits, visés dans des directives européennes. L'hypothèse d'une directive faisant des dispositifs d'intelligence artificielle des produits ou service nécessitant un marquage CE est évoquée implicitement dans le livre blanc de la commission européenne sur l'intelligence artificielle comme une piste à étudier[25].

Certains systèmes d'intelligence artificielle du fait de leur rattachement sectoriel à une de ces directives sont soumis à des évaluations pour permettre une apposition légitime du marquage CE, c'est le cas tout particulièrement des dispositifs médicaux utilisant l'intelligence artificielle. Dans ce contexte, par exemple en France, la Haute Autorité de Santé (ci-après « HAS ») a déployé un « Projet de grille d'analyse pour l'évaluation de dispositifs médicaux avec intelligence artificielle »[26], pour l'intégration des dispositifs médicaux sur la liste des produits et prestations remboursables. Selon la HAS, une telle évaluation devrait être réalisée après le marquage CE, ce qui peut s'expliquer par le fait que la HAS opère son évaluation en aval. Pour autant, une évaluation concomitante ou préalable pourrait être envisagée, cet enjeu paraissant plutôt en amont d'un marquage CE et d'une demande pour figurer sur une liste de produits remboursables.

Le marquage CE paraît donc largement dissocié de ce qui pourrait être pensé pour une certification de l'intelligence artificielle et l'Union européenne ne semble pas avoir entamé une réflexion approfondie sur une directive susceptible de créer une obligation de marquage CE spécifique à des systèmes d'intelligence artificielle et encore moins d'intelligence artificielle dans le domaine juridique, domaine dans lequel aucun produit ou service n'est visé par une directive européenne liée au marquage CE.

I.6 Mise en perspective de l'Ecolabel pour des systèmes d'intelligence artificielle

L'Ecolabel européen a été créé par le Règlement (CEE) n° 880/92 du Conseil, du 23 mars 1992, concernant un système communautaire d'attribution de label écologique puis révisé par le Règlement (CE) nº 1980/2000 du Parlement européen et du Conseil du 17 juillet 2000 établissant un système communautaire révisé d'attribution du label écologique. Ce label volontaire permet d'attester d'une qualité environnementale au cours du cycle de vie du produit par le respect d'un cahier des charges spécifique à chaque produit[27]. Pour obtenir l'Ecolabel, le produit doit être soumis à une procédure de certification par un organisme certificateur et le demandeur devra payer les frais de dossier puis des redevances[28].
La principale proximité entre l'Ecolabel et la certification de l'intelligence artificielle en matière de justice dont la faisabilité est discutée dans la présente étude est leur objectif sociétal à portée universaliste.
L'approche de l'Ecolabel repose principalement sur la comparaison et la valorisation des produits les plus respectueux des objectifs écologiques[29] et tient compte de l'évolution de la technique pour redéfinir régulièrement les critères[30]. Une certification d'aspects éthiques de l'intelligence artificielle pourrait s'inspirer de cette approche comparatiste et évolutive.

---

[23] Eric Lachaud, « ISO/IEC 27701: Threats and Opportunities for GDPR Certification », 2020, 24 pages.
[24] Communication de la Commission européenne, Le Guide bleu relatif à la mise en œuvre de la réglementation de l'Union européenne sur les produits 2016 (Texte présentant de l'intérêt pour l'EEE) (2016/C 272/01).
[25] Livre Blanc, Intelligence artificielle Une approche européenne axée sur l'excellence et la confiance, COM(2020) 65 final.
[26] https://www.has-sante.fr/jcms/p_3118247/fr/projet-de-grille-d-analyse-pour-l-evaluation-de-dispositifs-medicaux-avec-intelligence-artificielle
[27] Article 6, Règlement (CE) nº 1980/2000 du Parlement européen et du Conseil du 17 juillet 2000 établissant un système communautaire révisé d'attribution du label écologique.
[28] *Ibid.*, Article 12.
[29] *Ibid.* Considérant 6 : « Il convient d'expliquer au consommateur que le label écologique correspond à des produits qui sont susceptibles de réduire certains impacts négatifs sur l'environnement par comparaison avec d'autres produits de la même catégorie, sans préjudice des prescriptions réglementaires qui s'appliquent aux produits au niveau communautaire ou national. ».
[30] *Ibid.* Article 4.4 et 16.

I.7 Standards nationaux en élaboration pour l'intelligence artificielle : Malte et Danemark

Malte a entamé une réflexion approfondie sur la réglementation et la certification de l'intelligence artificielle, par la réalisation d'un document de consultation sur les aspects éthiques de l'intelligence artificielle[31] en août 2019 puis d'un document stratégique en octobre 2019 visant à faire de Malte la « base de lancement ultime de l'intelligence artificielle » (*Ultimate AI Launchpad)*[32].

L'objet de la consultation sur l'éthique de l'intelligence artificielle vise à aboutir à un code d'éthique pour l'intelligence artificielle permettant notamment un respect des droits fondamentaux. Ce code éthique dont la réalisation n'a pas encore abouti présente des proximités avec la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ mais son objet ne se limite pas au domaine juridique et judiciaire et son approche vise à être compatible avec la stratégie de compétitivité de Malte dans ce secteur. Le dernier chapitre du document de consultation est dédié à la certification de l'intelligence artificielle mais reste très allusif, l'objectif étant de créer de la confiance et de la transparence entre les utilisateurs, les consommateurs et les parties prenantes[33].

Le document stratégique propose d'implémenter des projets pilotes d'intelligence artificielle au sein du gouvernement et des administrations dans différents secteurs (gestion de la circulation, éducation, santé, service consommateur, tourisme). Le domaine de la justice n'a pas été inclus parmi ces domaines d'expérimentation[34]. La seule référence à l'intelligence artificielle en matière de justice dans ce document correspond au résultat d'un sondage selon lequel uniquement 50% des sondés seraient rassurés par une intelligence en matière de justice contre 80% dans le domaine des transports[35], ce qui semble avoir justifié son exclusion du champ de l'expérimentation.

Le Danemark fait également figure de pays actif dans la recherche d'une labellisation de l'intelligence artificielle en fonction de critères éthiques. Une stratégie nationale d'intelligence artificielle a été définie dans un document gouvernemental de mars 2019 visant la création d'une boite à outils à destination des entreprises et d'une labellisation en matière d'éthique de l'intelligence artificielle[36]. La création prochaine d'une labellisation sur la cybersécurité et l'usage responsable des données a été annoncée fin 2019 par le Ministère de l'Industrie du Danemark[37].

Ces stratégies nationales de labélisation saluées notamment par l'Union européenne[38] et l'OCDE[39] restent encore à l'état de projet en matière de certification éthique de l'intelligence artificielle et ne permettent pas de dégager des bases de travail ou de comparaison pour la présente étude. Néanmoins, ces dynamiques nationales démontrent la volonté pour des Etats membres du Conseil de l'Europe de disposer de certifications dans ce domaine sensible.

I.8 Les travaux de l'IEEE en matière de normalisation éthique de l'intelligence artificielle

L'Institut des ingénieurs électriciens et électroniciens (IEEE) est une organisation à but non lucratif de droit américain très active dans les normes techniques liées à l'électronique et à l'informatique qui se décrit comme la plus grande association dans le monde de professionnels pour les technologies avancées.

Au-delà de son activité de normes technique, l'IEEE affiche déjà un travail avancé prenant la forme d'une série d'actions intitulées P7000™ pour normaliser le « futur de l'éthique pour les agents autonomes et les systèmes intelligents ». Ces projets sont actuellement numérotés jusqu'à 14 (en sautant le n°13) :

- IEEE P7000™ : Model Process for Addressing Ethical Concerns During System Design
- IEEE P7001™ : Transparency of Autonomous Systems

---

[31] 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, août 2019, 36 pages.
[32] Parliamentary Secretariat For Financial Services Digital Economy and Innovation, Office of the Prime Minister, Malta : the Ultimate AI Launchpad, octobre 2019, 57 pages.
[33] 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, août 2019, pp. 33-34.
[34] [34] Parliamentary Secretariat For Financial Services Digital Economy and Innovation, Office of the Prime Minister, Malta : the Ultimate AI Launchpad, octobre 2019, p.2.
[35] *Ibid.*, p. 38.
[36] Danish Government, National Strategy for Artificial Intelligence, mars 2019, 74 pages : https://eng.em.dk/media/13081/305755-gb-version_4k.pdf
[37] https://eng.em.dk/news/2019/oktober/new-seal-for-it-security-and-responsible-data-use-is-in-its-way/
[38] Livre Blanc, Intelligence artificielle Une approche européenne axée sur l'excellence et la confiance, COM(2020) 65 final, p. 13.
[39] https://oecd.ai/wonk/an-independent-council-and-seal-of-approval-among-denmarks-measures-to-promote-the-ethical-use-of-data

- 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

Parmi ces projets de normes actuellement, seul le projet *IEEE 7010™-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being* a déjà abouti à une norme.

Parmi les projets listés ci-dessus, seuls les projets de normalisation P7000 à P7003 s'inscrivent réellement dans le champ de la présente étude sur la certification de l'intelligence artificielle dans le domaine judiciaire. Une ébauche du projet de normalisation P7000 peut déjà être commandée sur le site de l'IEEE mais les projets P7001 à 7003 n'ont pas encore donné lieu à la publication d'ébauches. Ces projets de normes P700X de l'IEEE ne concernent pas directement le domaine judiciaire pour l'instant et se réfèrent plus à l'éthique ou au bien être qu'au droits et libertés fondamentaux. Les questions de transparence (P7001™) et de biais algorithmiques (P7003™) sont essentielles et devront être prises en compte lorsque ces normes de l'IEEE seront effectivement élaborées et pourront permettre d'affiner les indicateurs utiles à la mise en œuvre du Principe n°4 de la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ et des critères de certification (*voir ci-dessous II*).

I.9 Inscription dans l'activité de labellisation du Conseil de l'Europe (Pharmacopée européenne et Label Itinéraire Culturel)

La pharmacopée européenne créée par des Etats membres sous l'égide du Conseil de l'Europe par convention en 1964[40] s'est progressivement ouverte à des acteurs de la normalisation. La pharmacopée est administrée par le Conseil de l'Europe via sa Direction européenne de la qualité du médicament et soins de santé (EDQM) qui emploie actuellement près de 400 personnes de 25 nationalités différentes. L'EDQM a signé un protocole d'accord avec le Comité européen de normalisation (CEN) concernant les dispositifs médicaux[41].

La fiche d'information de l'EDQM indique que la pharmacopée européenne « est la référence officielle utilisée par tous les professionnels impliqués dans la fabrication et le contrôle des médicaments. Elle a pour objectif de définir des exigences relatives aux médicaments et à leurs composants, qui soient juridiquement contraignantes »[42]. La conformité aux exigences de la pharmacopée conditionne la mise sur le marché des médicaments dans 39 Etats membres. Le caractère contraignant de cette certification permet de donner une réelle assise aux préconisations et critères définis par l'EDQM.

Le rattachement de l'EDQM au Conseil de l'Europe est expliqué par le fait que « la mission de l'EDQM est d'œuvrer pour le droit humain fondamental que constitue l'accès à des médicaments et soins de santé de qualité, et de contribuer à la promotion et la protection de la santé humaine et animale ». L'objectif de la certification discutée dans la présente étude paraît encore plus immédiatement rattaché aux missions du Conseil de l'Europe en matière de défense des droits de l'Homme, en ce qu'elle vise à les adapter à l'univers numérique, tout particulièrement aux intelligences artificielles, dans le domaine judiciaire.

L'expérience de la pharmacopée européenne, bien que réalisée dans un champ très différent du domaine juridique et judiciaire peut être très instructive pour penser une certification contraignante préalable à la mise sur le marché et pour structurer une certification d'un point de vue institutionnel, organisationnel (*voir ci-dessous IV*) et dans la relation avec des organismes de normalisation et l'Union européenne.

[40] Convention relative à l'élaboration d'une Pharmacopée européenne de 1964, Strasbourg, 22.VII.1964.
[41] https://www.edqm.eu/fr/Historique-1964-1997.html
[42] https://www.edqm.eu/sites/default/files/medias/fichiers/AboutUs/edqm_fiche_dinformation_la_direction_europeenne_de_la_qualite_du_medicament_soins_de_sante.pdf

Le Conseil de l'Europe administre également le Label Itinéraire Culturel. Un retour d'expérience pourrait être utile pour la conception et le déploiement d'une certification du Conseil de l'Europe sur l'intelligence artificielle, d'un point de vue organisationnel (*voir ci-dessous IV*) que ce soit pour définir les cycles d'évaluation ou la création d'un institut dédié à la certification. Les équipes du Conseil de l'Europe impliquées dans la mise en œuvre du Label Itinéraire Culturel pourraient être sollicitées pour donner des indications sur les spécificités d'une labellisation à l'échelle du Conseil de l'Europe.

Cette expérience du Label Itinéraire Culturel pourrait être instructive pour apprécier les convergences et coopérations possibles avec d'autres institutions internationales ayant un rôle de labellisation culturel et ayant également des travaux de fond sur l'intelligence artificielle, tout particulièrement l' Organisation des Nations unies pour l'éducation, la science et la culture (ci-après « UNESCO ») et également dans une certaine mesure l' Organisation de coopération et de développement économiques (ci-après « OCDE »).

I.10 Analyse d'impact type RGPD et certification

Les études d'impact réalisées par des entreprises ont été systématisées avec le RGPD qui impose dans ses articles 35 et 36, ce type de documents lorsqu'un traitement de données personnelles est « susceptible d'engendrer un risque élevé pour les droits et libertés des personnes physiques, le responsable du traitement effectue, avant le traitement, une analyse de l'impact des opérations de traitement envisagées sur la protection des données à caractère personnel »[43].

Une analyse d'impact pour les intelligences artificielles pourrait être envisagée sur le modèle de celle du RGPD, en tenant compte non seulement des conséquences sur la protection des données personnelles mais surtout des incidences sociétales et économiques de décisions inexplicables ou biaisées prises par une intelligence artificielle.

I.11 Bac à sable d'intelligence artificielle contextualisée et certification

Le terme bac à sable est couramment utilisé dans le domaine de l'informatique dans le cadre de tests de programmes avant leur déploiement, notamment pour identifier les bugs et assurer un lancement dans de bonnes condition. Il s'agit principalement d'une pratique de sécurité informatique permettant d'exécuter un programme dans un environnement restreint[44].

Cette pratique informatique a été récemment transposée dans l'univers juridique avec les bacs à sable réglementaires, initialement dans le domaine de la régulation financière. Le projet pionner en la matière intitulé *Project Innovate* est un programme porté depuis 2014 par une instance britannique de régulation financière, la *Financial Conduct Authority* (ci-après « FCA »). La FCA indique que cette pratique permet à des projets commerciaux d'expérimenter des propositions innovantes sur le marché avec de vrais consommateurs et indique que cette approche permet de tester un produit ou service dans un environnement contrôlé, de réduire le temps de mise sur le marché, d'aider à identifier les précautions à intégrer à l'égard du consommateur avant la mise sur le marché et un meilleur financement du projet[45]. Cette pratique a largement essaimé depuis 2014, elle est globalement plébiscitée dans la Fintech et également dans l'intelligence artificielle[46], notamment en France[47] ou en Finlande[48].

Dans la perspective d'une certification d'intelligence artificielle en matière judiciaire du Conseil de l'Europe, un des bénéfices d'un bac à sable serait la sensibilisation des acteurs aux enjeux de droits de l'Homme dès la conception et d'éthique. Un bac à sable permettrait sur une période prolongée d'accompagner les acteurs dans leur appréhension du contenu de la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ et des critères de certification.

---

[43] Article 35, Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données).
[44] VassilisPrevelakis et Diomidis Spinellis, « Sandboxing Application », Proceedings of the USENIX Technical Annual Conference, Freenix Track, June 2001, pp. 119–126 (2001).
[45] https://www.fca.org.uk/firms/innovation/regulatory-sandbox
[46] Voir plus généralement : Laura Delponte, « Study : European Artificial Intelligence (AI) leadership, the path for an integrated vision », European Parliament, IPOL, 2018, 48 pages.
[47] Cédric Villani, Donner un sens à l'intelligence artificielle (IA), Rapport, Mission parlementaire française, 2018, 233 pages.
[48] 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/

Le bac à sable permettrait également d'affiner et d'améliorer constamment les critères de certification au plus près de la pratique et de l'évolution de la technique.

Pour les porteurs de projets d'intelligence artificielle dans le domaine judiciaire, au-delà de la certification susceptible d'être obtenue au cours ou à l'issue d'une période de bac à sable, cet accompagnement pourrait permettre d'améliorer les chances de succès commercial de leur dispositif, leurs clients (Etats, administrations, professionnels réglementés du droit…) étant particulièrement attentifs au respect de la réglementation et à l'effort de mise en conformité. Concernant les projets directement portés par des Etats ou administrations, un bac à sable et d'une manière générale une certification dans ce domaine serait un accompagnement utile permettant de s'aligner sur des standards supranationaux et de réduire les risques d'une condamnation par la Cour Européenne des Droits de l'Homme (ci-après « CEDH ») dans un domaine stratégique.

Néanmoins, comme évoqué précédemment, les intelligences artificielles connexionnistes étant évolutives, la certification attribuée, par exemple, à l'issue d'une période de bac à sable ne peut être conservée de manière définitive.

Si actuellement aucune certification de l'intelligence artificielle en matière judiciaire ne peut être utilisée comme source d'inspiration, des réflexions ont été entamées en matière de certification d'intelligence artificielle et plus globalement des invariants en matière de certifications permettent de concevoir un tel dispositif. L'expérience du Conseil de l'Europe en matière de labélisation et de certification avec les itinéraires culturels et la pharmacopée européenne peut s'avérer utile pour structurer institutionnellement la certification envisagée. Par ailleurs, cette certification pourrait être un moyen de diffuser et de faire appliquer effectivement la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ grâce à des méthodes d'accompagnement comme les bacs à sable.

II. Formalisation des critères et indicateurs d'une certification correspondant à la Charte

II.1 Méthode de formalisation d'indicateurs à partir des principes de la Charte

Afin de retranscrire fidèlement dans une certification la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ, il convient de concevoir des indicateurs opérationnels, objectifs et vérifiables en s'appuyant sur les cinq principes qui en sont issus :

- Principe de respect des droits fondamentaux : assurer une conception et une mise en œuvre des outils et des services d'intelligence artificielle qui soient compatibles avec les droits fondamentaux.
- Principe de non-discrimination : prévenir spécifiquement la création ou le renforcement de discriminations entre individus ou groupes d'individus.
- Principe de qualité et sécurité : en ce qui concerne le traitement des décisions juridictionnelles et des données judiciaires, utiliser des sources certifiées et des données intangibles avec des modèles conçus d'une manière multidisciplinaire, dans un environnement technologique sécurisé.
- Principe de transparence, de neutralité et d'intégrité intellectuelle : rendre accessibles et compréhensibles les méthodologies de traitement des données, autoriser les audits externes.
- Principe de maîtrise par l'utilisateur : bannir une approche prescriptive et permettre à l'usager d'être un acteur éclairé et maître de ses choix.

II.2 Indicateurs pour le principe n°1 : respect des droits fondamentaux

Principe n°1 de la Charte : « Principe de respect des droits fondamentaux : assurer une conception et une mise en œuvre des outils et des services d'intelligence artificielle qui soient compatibles avec les droits fondamentaux. »

Le premier principe relatif au respect des droits fondamentaux est le plus spécifique et certainement le plus complexe à transcrire dans une certification de l'intelligence artificielle. Ce premier principe s'appuie notamment sur le concept de droits de l'Homme dès la conception (*Human rights by design*) qui fait figure de ligne directrice de la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ.

- Traitement proportionné des données personnelles (vie privée) et finalités claires

- Anonymisation des parties et intervenants personnes physiques et de leurs conseils
    - Vérification par consultation des jeux de données
- Absence de notations et classements de personnes physiques ou morales sur la base de décisions de justice
    - Vérification par consultation de l'interface

- Anonymisation du nom du juge et de la localisation de la juridiction dans les décisions utilisées pour de la justice prédictive (objectif éviter le *forum shopping*)
    - Vérification par consultation des jeux de données
- Etanchéité entre les services d'intelligence artificielle ayant des finalités différentes (par exemple : dissocier le service de moteur de recherche du service d'aide à la décision)
    - Vérification des bases de données et des sources de données utilisées par chaque système

- Droit d'accès au juge et droit au procès équitable

- Présence d'une mention lisible indiquant, le cas échéant, le caractère non explicable d'un rapport issu d'une intelligence artificielle. Cet aspect permet au juge de garder la pleine maîtrise dans sa prise de décision et de n'utiliser les rapports non entièrement explicables, susceptibles d'être générés par des intelligences artificielles connexionnistes, qu'en connaissance de cause.
    - Modèle d'apprentissage et interface
- Pour les aspects liés à l'accès au juge se reporter aux indicateurs liés à la maîtrise par l'utilisateur et plus spécifiquement à la possibilité d'opposition par le justiciable au recours à l'IA (II.6)

- Indépendance des juges dans leur processus de décision

- Garantie contre le profilage des juges
    - Vérification A/B testing sur moteur de recherche avec des comptes utilisateurs différents et des historiques de recherche différents
- Adéquation entre le critère affiché et le mode de classement effectif des résultats de recherche
    - Vérification par audit des résultats de recherche
- Transparence des pondérations entre critères en cas de recherches multicritères
    - Vérification de l'existence de mentions explicatives et audit des résultats de recherche
- Transparence des critères utilisés en cas de recherche dite par « pertinence »
    - Vérification de l'existence de mentions explicatives et audit des résultats de recherche

- Ethique et droits de l'Homme dès la conception

- Prise en compte *ab initio* des droits et libertés fondamentaux et de la CESDH :
    - Pour les intelligences artificielles symboliques
        - Rapport de présentation des arbres de décision expliquant la prise en compte des droits et libertés fondamentaux
    - Pour les intelligences artificielles connexionnistes à entraînement supervisé
        - Rapport de présentation des données d'entraînement et des modalités d'entraînement expliquant la prise en compte des droits et libertés fondamentaux
        - Analyse d'impact du service utilisant l'intelligence artificielle en fonction des droits et libertés fondamentaux : Absence de vérification *a priori* (responsabilisation des acteurs)
        - Désignation facultative d'un délégué à l'intelligence artificielle indépendant en cas d'explicabilité suffisante et obligatoire en cas d'explicabilité insuffisante
    - Pour les intelligences artificielles connexionnistes à entraînement non-supervisé
        - Analyse d'impact du service utilisant l'intelligence artificielle en fonction des droits et libertés fondamentaux : Absence de vérification *a priori* (responsabilisation des acteurs)
        - Désignation obligatoire d'un délégué à l'intelligence artificielle indépendant
- Installation et permission d'accès continue à un plugin CEPEJ de conformité aux droits de l'Homme (méta-intelligence artificielle/plugin susceptible d'être développé ultérieurement grâce aux informations obtenues à partir des analyses d'impact, un tel plugin pourrait potentiellement être utilisé ou réadapté au-delà du champ juridique et judiciaire).

II.3 Indicateurs pour le principe n°2 : non-discrimination

Principe n°2 de la Charte : « Principe de non-discrimination : prévenir spécifiquement la création ou le renforcement de discriminations entre individus ou groupes d'individus »

Le deuxième principe vise à éviter la discrimination en fonction de données dites sensibles. L'intelligence artificielle repose par nature sur la création de catégories plus ou moins fines, donc sur la discrimination. Limiter la discrimination revient à brider l'intelligence artificielle et supprimer des données sensibles peut créer des biais, qui peuvent être préférables ou nécessaires. L'anonymisation ou l'exclusion sélective de certaines données permet de limiter la discrimination[49]. Certaines recherches, comme celle issue de chercheurs de Eindhoven University of Technology proposent des solutions de retraitement des données pour supprimer des

[49] Sur cette question voir par exemple : Salvatore Ruggieri, « Data Anonymity Meets Non-Discrimination », IEEE 13th International Conference on Data Mining Workshops (December 7-10, 2013), pp. 875–88.

effets discriminatoires[50], mais celles-ci paraissent très exigeantes et difficiles, pour l'instant, à déployer au-delà d'un cadre expérimental. Certaines informations sensibles, par exemple, celles liées à l'état de santé peuvent avoir pour effet la déresponsabilisation ou constituer des circonstances atténuantes. Néanmoins, pour éviter effectivement une discrimination sur des fondements contraires aux droits de l'Homme et ne pas amplifier des inégalités, il convient d'appréhender de manière extensive ces données sensibles en y incluant des marqueurs susceptibles de produire indirectement les mêmes effets discriminatoires par recoupement (déduction de l'origine à partir du nom de famille, déduction de l'orientation sexuelle à partir d'éléments de contexte, déduction du milieu socio-économique à partir de l'adresse du domicile…). Il convient également de noter que l'utilisation de l'intelligence artificielle est susceptible de faire émerger de nouvelles formes de discrimination très profilées et non catégorielles, moins immédiatement lisibles, par exemple, qu'une discrimination fondée sur l'ethnie.

Ce principe doit être apprécié en parallèle des prescriptions de l'article 6 de la Convention 108+ concernant les catégories particulières de données, et des articles 9 et 10 du RGPD également applicables dans de nombreux Etats membres du Conseil de l'Europe qui n'autorisent le traitement de ce type de données qu'à la condition que des garanties appropriées soient fournies en complément des obligations s'appliquant aux données personnelles classiques.

L'article 9 du RGPD précise notamment comme exceptions le traitement manifestement nécessaire à la constatation, à l'exercice ou à la défense d'un droit en justice et le traitement nécessaire pour des motifs d'intérêt public important, sur la base du droit de l'Union ou du droit d'un Etat membre qui doit être proportionné à l'objectif poursuivi.

L'article 10 du RGPD avance une distinction en fonction de l'entité réalisant le traitement pour les données personnelles relatives aux condamnations pénales et aux infractions ou aux mesures de sûreté, celles-ci ne pouvant être traitées que sous le contrôle de l'autorité publique et dans des cas exceptionnels et avec les garanties appropriées par d'autres entités.

L'article 6 de la Convention 108+ vise explicitement le risque de discrimination résultant du traitement des données sensibles qui se retrouve dans la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ.

- Non-discrimination fondée sur des données sensibles

Par principe :

- Suppression des marqueurs rattachables aux données sensibles des parties (domicile, revenu, situation familiale, capital social)
    - Vérification par consultation des jeux de données
    - A/B testing à partir d'information et marqueurs rattachables à des données sensibles en changeant, le cas échéant, lors de chaque test un des paramètres suivants : nom, domicile, revenu, situation familial, capital social, élément de contexte spécifique pertinent
    - Formulaire donnant la possibilité pour les utilisateurs d'exprimer de manière circonstanciée des demandes de suppression de marqueurs rattachables à des données sensibles avec copie au délégué à l'intelligence artificielle, le cas échéant, et à l'autorité de contrôle (entité attribuant la labellisation).

Par exception, pour des intelligences artificielles sous le contrôle de l'autorité publique et pour des données relatives aux condamnations pénales et aux infractions ou aux mesures de sûreté :

- Dispositifs garantissant le caractère nécessaire et proportionné du traitement
    - Analyse d'impact par l'organisme de traitement expliquant les garanties mises en œuvre
    - Désignation d'un délégué à l'intelligence artificielle

II.4 Indicateurs pour le principe n°3 : qualité et sécurité de la donnée

Principe n°3 de la Charte : « Principe de qualité et sécurité : en ce qui concerne le traitement des décisions juridictionnelles et des données judiciaires, utiliser des sources certifiées et des données intangibles avec des modèles conçus d'une manière multidisciplinaire, dans un environnement technologique sécurisé. »

La norme ISO/IEC 27001 précitée (voir ci-dessus I.5) fait figure en 2020 de norme de référence en matière de sécurité informatique et des données. La norme ISO/IEC 27001 prévoit une centaine de mesures de sécurité conçues pour vérifier la robustesse d'un système en matière de cybersécurité. Le respect de cette norme

[50] Voir par exemple Faisal Kamiran et Calder Toon, « Data Preprocessing Techniques for Classification Without Discrimination », Knowledge and Information Systems 33, no. 1 (December 3, 2011), pp. 1–33.

technique pourrait être prévu comme un préalable à une demande relative à la présente certification. En matière de cybersécurité, le recours à une norme technique ISO/IEC, conçue principalement par des entreprises du secteur, ne paraît pas poser de difficultés structurelles dans la mesure où en la matière, celles-ci cherchent à appliquer les standards les plus élevés pour atteindre des obligations de résultats vis-à-vis de leurs clients et respecter des lois et réglementations contraignantes en la matière. Lorsqu'une certification au sens de l'article 42 du RGPD sera adoptée en matière de cybersécurité, celle-ci pourra être envisagée pour être substituée à la norme ISO/IEC.

La qualité de la donnée et de sa non-altération sémantique[51] jusqu'à l'intégration dans le jeu de données mis à disposition de l'intelligence artificielle, est un point fondamental. Cet enjeu est régulièrement désigné sous le terme ALCOA[52] et ALCOA+[53], les données devant être attribuables, lisibles, à jour, d'origine, conformes à la réalité, complètes, cohérentes, durables et accessibles.

Une méthode envisageable pour s'assurer de l'intégrité des « données dérivant des décisions juridictionnelles » visées dans le troisième principe de la Charte pourrait être de créer un système permettant aux autorités compétentes des pays membres de procéder à des ancrages sur blockchain des décisions de justices numérisées. Pour autant les certifications, tout particulièrement quand elles sont obligatoires peuvent dans certains cas créer des barrières indirectes à l'entrée du marché.

Ce troisième principe évoque également le caractère multidisciplinaire de la conception ou *a minima* de l'évaluation des modèles d'intelligence artificielle. Un tel objectif pourrait être atteint par la création d'un groupe d'experts composé de professionnels du droit, de chercheurs et d'enseignants chercheurs issus de disciplines telles que notamment, le droit, les sciences de l'information et de la communication, les mathématiques, l'informatique, l'économie ou encore la sociologie. Ce groupe pourrait structurer scientifiquement un dispositif d'accompagnement des porteurs de solutions d'intelligence artificielle dans le domaine juridique et judiciaire, particulièrement dans le cadre de bacs à sable (voir ci-dessus I.11).

II.5 Indicateurs pour le principe n°4 : transparence, neutralité et intégrité intellectuelle

Principe n°4 de la Charte : « Principe de transparence, de neutralité et d'intégrité intellectuelle : rendre accessibles et compréhensibles les méthodologies de traitement des données, autoriser les audits externes. »

- Accès au code source pour le besoin de la certification

Les références figurant dans la Charte mettant en perspective ce principe, notamment l'étude MSI-NET du Conseil de l'Europe, le rapport Villani et le rapport de la House of Lords, affichent une certaine résignation quant à la possibilité d'obtenir la divulgation publique d'algorithmes entiers ou de leur code source et envisagent uniquement le partage d'informations partielles mais pertinentes[54]. Dans l'hypothèse d'une certification facultative (voir I.1), prévoyant, comme condition à l'obtention de la certification, la divulgation totale du code source, il est possible que les acteurs disposant des algorithmes les plus avancés ne souhaitent pas se priver d'avantages concurrentiels même en contrepartie d'une certification susceptible de rassurer leurs clients.

Dans l'hypothèse d'une certification obligatoire, l'obligation associée de divulgation du code source paraît pleinement opératoire, mais devrait être mise en œuvre en respectant au maximum les secrets d'affaires.

---

[51] L'altération sémantique n'inclut pas par exemple l'exclusion de données sensibles, telles que celles dont la suppression a été envisagée précédemment (voir. II.3).

[52] Acronyme anglais pour : Attributable ; Legible ; Contemporaneous ; Original ; Accurate.

[53] Le « + » d'ALCOA+ vise les caractéristiques suivantes : Complete ; Consistent ; Enduring ; Available.

[54] Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ, page 11, note n°3 « La proposition faite dans l'étude du MSI-NET du Conseil de l'Europe intitulée «Algorithmes et droits humains», page 46, est intéressante: «La solution consistant à demander la divulgation publique d'algorithmes entiers ou de leur code source est utopique dans ce contexte, les entreprises privées considérant leurs algorithmes comme des logiciels propriétaires stratégiques, et les protégeant donc en conséquence. Il paraît en revanche envisageable d'exiger la publication d'informations partielles mais néanmoins importantes, comme les variables utilisées, les objectifs visés par l'optimisation des algorithmes, les jeux de données d'apprentissage, les valeurs moyennes et les écarts types des résultats obtenus, ou la quantité et le type de données traitées par l'algorithme ». Ou encore les suggestions du rapport intitulé «l'IA pour l'humanité» rédigé par le député Cédric Villani dans le cadre d'une mission confiée par le Premier Ministre de la République française, à la page 144: «les auditeurs pourraient tester l'équité et la loyauté d'un programme (faire ce qui est censé faire), par exemple à travers la soumission de multiples fausses données d'entrée, créer des nombreux profils d'utilisateur suivant des caractéristiques précises»...Mêmes constats dans le rapport de la House of Lords, «AI dans le Royaume Uni: prêts, disposés et capables? sur l' IA: paras 92, 96-99) ».

Il serait ainsi possible d'envisager une transmission des codes sources complets sous couvert de confidentialité au comité attribuant la certification ou au délégué à l'intelligence artificielle, éventuellement après homologation *intuitu personae* par l'acteur concerné.

En outre, concernant les intelligences artificielles contrôlées par des Etats, celles conçues en interne par des autorités publiques devraient être *open source*, au même titre que les textes de loi sont accessibles aux justiciables et celles conçues par des prestataires et utilisées par des Etats pourraient n'être sélectionnées qu'à la condition d'être *open source*.

L'accès au code source n'est pas suffisant pour assurer une pleine explicabilité de certaines intelligences artificielles.

- FAT-ML : Intelligence artificielle équitable, responsable et transparente

Un certain consensus pour l'étude du comportement des algorithmes met l'accent sur le caractère équitable, responsable et transparent de l'intelligence artificielle, cette approche est dite FAT-ML (Fairness, Accountability, Transparency in Machine Learning)[55].

- **Caractère équitable**

La question du caractère équitable de l'intelligence artificielle recoupe celle du caractère non discriminatoire visée dans troisième principe de la Charte (voir II.4).

- **Responsabilité du fait de l'intelligence artificielle**

Le critère de responsabilité fait référence à la possibilité d'engager la responsabilité d'une personne physique ou morale en cas de dysfonctionnement ou de dommage causé par l'intelligence artificielle. La certification pourrait potentiellement avoir pour effet d'exonérer au moins partiellement de sa responsabilité l'entité de bonne foi en cas de problème lié à l'intelligence artificielle certifiée. La logique de bac à sable déjà évoquée a pour objectif de limiter le dommage et donc la responsabilité susceptible d'en découler, grâce à une première mise en service contrôlée et restreinte pendant une période d'expérimentation.

- **Transparence**

Au-delà de l'accès au code, la transparence désigne la possibilité d'expliquer les décisions de l'intelligence artificielle. Ce caractère de l'intelligence artificielle est essentiel et a déjà été largement pris en compte dans les indicateurs proposés pour le premier principe comme prérequis essentiel de tout dispositif d'éthique et de droit de l'Homme par conception en distinguant en fonction du type d'intelligence artificielle, symbolique ou connexionniste (voir II.2).

Les intelligences artificielles symboliques sont déterministes, figées et explicables puisque retranscrivant un mode de raisonnement humain. La divulgation du modèle suffit donc à auditer l'intelligence artificielle.

Les intelligences artificielles connexionnistes peuvent devenir difficiles à expliquer lorsqu'elles évoluent et se complexifient. Pour cette catégorie d'intelligence artificielle un contrôle continu en recourant à un délégué à l'intelligence artificielle ou à un plugin contrôlant automatiquement l'intelligence artificielle doit être envisagé pour viser l'explicabilité. En l'absence d'explicabilité, cet état doit *a minima* être immédiatement lisible par l'utilisateur accédant au résultat fourni par l'intelligence artificielle (voir I.2 et I.3).

Concernant la transparence, la norme IEEE P7001™ *Transparency of Autonomous Systems* en cours d'élaboration devrait permettre d'intégrer des spécifications techniques en vue de renforcer la protection de ce principe de transparence.

- **Identification de l'intelligence artificielle et de ses actions**

La question de l'identification est un point fondamental pour assurer la transparence et la responsabilité des acteurs du secteur. Au même titre qu'un utilisateur est indirectement identifiable dans sa navigation et ses communications, il paraît utile qu'une intelligence artificielle, ayant une certaine autonomie d'action le soit également et puisse être identifiable explicitement sur le réseau comme étant une intelligence artificielle (éventuellement en précisant la catégorie d'intelligence artificielle, connexionniste ou symbolique). Actuellement un utilisateur est identifiable indirectement sur un réseau via l'adresse MAC de ses périphériques ou dans une certaine mesure sur Internet via l'adresse IP du matériel de connexion lorsque celle-ci est fixe. Des outils d'identification renforcée des utilisateurs se développent lorsque leur action nécessite une identification certaine (dispositifs de KYC, signature électronique…).

Cette identification de l'intelligence artificielle permet que les collectes opérées directement par celle-ci puissent être connues (que ses traces soient lisibles) et que ses actions puissent lui être attribuées. Cette

[55] Voir notamment le site : https://www.fatml.org/

dimension est particulièrement indispensable pour les systèmes complexes avec des intelligences artificielles composites, qui peuvent avoir été développées par des entités différentes.

L'identification ou l'immatriculation, régulièrement mises en avant pour les objets connectés sont globalement exclues des discussions pour l'intelligence artificielle en raison des contraintes potentielles pour les industriels et des usages dans le domaine de la défense, mais ce point est structurel pour l'effectivité d'une norme ou d'une certification fondée sur l'éthique et la responsabilisation des acteurs.

Un critère de certification pourrait être l'immatriculation de l'intelligence artificielle et l'obligation de signature par l'intelligence artificielle de ses actions. Un registre d'intelligences artificielles certifiées pourrait ainsi être créé et une signature unique leur être attribuée, sur le modèle des adresses IP attribuées par l'IANA de l'ICANN pour les sites internet et des outils de type signature de code pourraient être adaptés aux intelligences artificielles. L'utilisation d'une blockchain pourrait amplifier la transparence d'un tel registre et renforcer sa sécurité, renforçant ainsi la confiance des utilisateurs.

II.6 Indicateurs pour le principe n°5 : maîtrise par l'utilisateur

Principe n°5 de la Charte : « Principe de maîtrise par l'utilisateur : bannir une approche prescriptive et permettre à l'usager d'être un acteur éclairé et maître de ses choix. »

Le principe de maîtrise par les différents types d'utilisateurs est un corolaire de l'explicabilité auquel s'ajoute un enjeu d'interfaçage, particulièrement d'interface utilisateur (*user interface* souvent abrégée en UI) et d'expérience utilisateur (*user experience* souvent abrégée en UX).

- Garantir à l'utilisateur la maîtrise du dispositif

La maîtrise du dispositif par les utilisateurs vise à limiter l'opacité du fonctionnement de la plateforme et à permettre à l'utilisateur de comprendre et utiliser l'ensemble des fonctionnalités disponibles.

Cette maîtrise peut être facilitée par :

- une assistance téléphonique disponible et pertinente
- une interface ergonomique et intuitive
- des guides d'utilisation clairs et détaillés,
- des tutoriels courts axés sur la prise en main de chaque fonctionnalité
- des formations à l'outil

Dans le cadre d'une certification, par exemple à l'issue de la phase de bac à sable, en fonction de la complexité de la plateforme et de son interface, des mesures à prendre pour atteindre cet objectif de maîtrise par l'utilisateur pourraient être préconisées par l'organisme certificateur en vue de l'obtention de la certification. Pour s'assurer de l'efficacité des dispositifs d'accompagnement, des tests auprès d'utilisateurs « moyens » pourraient être réalisés pour en mesurer l'efficacité.

- Opposition du justiciable (AI opt-out)

La Charte préconise d'informer clairement le justiciable en cas d'utilisation d'une intelligence artificielle pour rendre une décision de justice et de prévoir la possibilité de s'y opposer et de demander à ce que son affaire soit entendue directement par un tribunal au sens de l'Article 6 de la CEDH.

- Information lisible pour le justiciable de l'utilisation d'une intelligence artificielle pour son affaire
    - Vérification de la présence d'une bannière d'information mentionnant l'utilisation d'une intelligence artificielle devant être cliquée avant tout accès au service
- Droit d'opposition du justiciable à l'utilisation d'une intelligence artificielle
    - Vérification de la présence d'un module d'expression du consentement à la dernière ligne du déroulé de la bannière d'information
    - Vérification d'un système de notification de la décision du justiciable et de redirection effective vers une procédure classique devant un tribunal au sens de l'Article 6 de la CEDH

- Formation et certification des magistrats

Dans l'hypothèse d'une solution d'intelligence artificielle d'aide à la décision judiciaire déployée par un organisme étatique en vue de son expérimentation puis de sa généralisation, une certification des magistrats paraîtrait également utile, en complément de la certification de la plateforme en tant que telle. Une telle formation certifiante pourrait être axée sur la compréhension des fonctionnalités et la méthode critique à adopter à l'égard des propositions de l'intelligence artificielle lors de l'élaboration d'une décision de justice assistée.

III. Autorités et modalités de certification

Les principales modalités de certification en fonction des autorités compétentes sont les suivantes :

- Auto-évaluation par les éditeurs de l'intelligence artificielle
- Evaluation par la CEPEJ ou un institut rattaché au Conseil de l'Europe
- Evaluation par des organismes notifiés
- Evaluation en continu par un plugin d'intelligence artificielle développée pour ou par la CEPEJ
- Evaluation mixte et évolutive

Le choix des modalités et des autorités de certification est en partie induit par d'autres caractéristiques de la certification. Le caractère obligatoire ou facultatif, le degré de contrôle souhaité, la rentabilité économique de le la procédure de certification ou à l'inverse le caractère déficitaire du dispositif, la complexité technique, l'existence d'organismes ayant des compétences éprouvées pour la certification de dispositifs numériques en fonction de critères éthiques ou encore la faisabilité d'une automatisation sont autant de critères susceptibles de guider le choix du type d'autorité de certification à privilégier.

III.1 Auto-évaluation par les éditeurs de l'intelligence artificielle

L'auto-évaluation par les éditeurs et/ou sous-traitants de l'intelligence artificielle a pour principal avantage l'effet de responsabilisation. Les aspects d'auto-évaluation discutés dans la présente étude s'inspirent principalement des modalités mises en œuvre par le RGPD en matière de données personnelles avec les analyses d'impact et le recours à des experts indépendants. Cette approche permet de conscientiser les éditeurs et d'inciter à une vigilance continue pour autant elle ne paraît pas suffisante dans un domaine aussi sensible que celui de l'intelligence artificielle en matière de justice, tout particulièrement pour des intelligences artificielles connexionnistes.

L'auto-évaluation, comme seule méthode utilisée, paraît difficilement compatible avec une certification, particulièrement avec une certification obligatoire.

III.2 Evaluation par la CEPEJ ou un institut rattaché au Conseil de l'Europe

L'évaluation par la CEPEJ ou un organisme rattaché au Conseil de l'Europe spécifiquement dédié à la gestion de cette certification pourrait être une option. Ce modèle pourrait s'inspirer du mode d'attribution du certificat de conformité à la pharmacopée européenne avec l'appui de la Direction européenne de la qualité du médicament et soins de santé (DEQM) du Conseil de l'Europe et la Commission européenne de la Pharmacopée créée à cet effet par le Comité de santé publique du Conseil de l'Europe.

Les modalités d'attribution de la certification de conformité aux monographies de la Pharmacopée européenne sont définies dans une résolution AP-CSP(07) du Comité de Santé Publique du Conseil de l'Europe du 21 février 2007. Ce document présente le champ d'application, la qualité du titulaire du certificat et la procédure de délivrance du certificat (dépôt de dossier, accusé réception, désignation des évaluateurs, les modalités d'évaluation, notification de la décision, le suivi de la certification de conformité et documents de référence).

Parmi les avantages d'une certification directe ou quasi-directe par la CEPEJ, il est possible de citer le cadre procédural propice à une application fidèle à l'esprit des principes de la Charte, le rayonnement supérieur de la certification, la gestion plus directe des effets potentiels d'un lobbying d'acteurs et industriels de l'intelligence artificielle ou encore la sélection directe des personnes physiques en charge des évaluations et le contrôle de leur impartialité, de leur intégrité et de leur compétence. Une certification directe ou quasi-directe peut également présenter une légitimité supérieure justifiant son caractère obligatoire.

Le déploiement d'une certification représente un coût de fonctionnement, qui est plus important si la procédure est gérée directement, néanmoins les frais de dossier exigibles pour la procédure de certification peuvent être définis de manière à couvrir les coûts de fonctionnement et même à dégager un bénéfice susceptible, par exemple, d'être affecté à l'amélioration du dispositif de cadrage de l'intelligence artificielle en matière judiciaire, au développement d'un plugin de contrôle des intelligences artificielles ou à l'adaptation de cette certification à d'autres secteurs dans lesquels l'intelligence artificielle peut porter atteinte aux droits de l'Homme, notamment en exacerbant des discriminations.

III.3 Evaluation par des organismes notifiés

Les organismes notifiés ou organismes certificateurs sont des organismes habilités spécifiquement pour appliquer de manière indépendante le cahier des charges et la procédure relative à une certification issue d'un organisme public.

Ces organismes notifiés ont une expertise importante en matière de conduite de certification. La mise en place de procédure de certification peut être plus rapide qu'une procédure conduite devant un organisme public créant une structure dédiée à cet effet. En outre, ces organismes inscrivent ces certifications dans une économie d'échelle permettant de limiter les coûts de ces procédures et de réaliser des bénéfices d'exploitation.

Les certifications entièrement déléguées à des organismes notifiés, lorsqu'elles sont simplement facultatives, peuvent perdre en attractivité pour les entités souhaitant obtenir pour le produit ou service une certification ayant une réelle reconnaissance.

III.4 Evaluation en continu par un plugin d'intelligence artificielle développé pour ou par la CEPEJ

Un plugin d'intelligence artificielle développé pour contrôler la conformité en continu des dispositifs utilisant de l'intelligence artificielle en matière judiciaire, déjà évoqué (ci-dessus I.3, II.2, II.5) pourrait permettre de prendre en charge automatiquement le suivi de l'évaluation, voire une partie de l'évaluation. Un tel plugin pourrait faciliter le travail d'un délégué à l'intelligence artificielle dans sa mission de suivi de conformité et éventuellement s'y substituer au moins partiellement. Un tel outil s'il était développé engendrerait des coûts de développement, de maintenance, un temps important avant de prouver son efficacité et un recul suffisant sur les difficultés rencontrées par des humains lors de la procédure et le suivi de la certification.

III.5 Evaluation mixte et évolutive

Une évaluation mixte et évolutive pourrait tirer profit des différentes caractéristiques évoquées pour chacune des modalités et autorités d'évaluation.

Une première auto-évaluation pourrait être ainsi réalisée par l'éditeur, demandeur de la certification, sous la forme des analyses d'impacts visées précédemment (voir II.) et le suivi de l'évaluation après certification pourrait en partie être réalisé par un délégué à l'intelligence artificielle désigné par l'éditeur.

Une option pragmatique pourrait également consister à déléguer une procédure de pré-certification ou l'évaluation de la partie technique (principalement évaluation de la sécurité des systèmes d'information et de la transparence) à ce type d'organismes notifiés afin, par exemple, de focaliser le travail d'évaluation de la CEPEJ ou d'un organisme lui étant rattaché sur les aspects strictement liés au droits de l'Homme dès la conception et de lutte contre les discriminations dans les solutions d'intelligence artificielle en matière judiciaire.

Le recours à un plugin d'intelligence artificielle pourrait également être un outil complémentaire utile pour limiter les contraintes liées au suivi de l'évaluation après la certification. Néanmoins, cet outil ne peut être envisagé que dans un second temps, celui-ci devant être conçu en s'appuyant sur des analyses d'impacts et des rapports de délégués à l'intelligence artificielle.

IV. Structure de gouvernance

IV.1 Modèles de gouvernance

Le modèle de gouvernance de la Direction européenne de la qualité du médicament et soins de santé pourrait être une source d'inspiration dans l'hypothèse de la création d'une équipe dédiée à la certification de l'intelligence artificielle en matière de justice dans le cas, par exemple, d'un accord partiel sur ce thème au sein du Conseil de l'Europe[56].

Un organigramme pour un service de certification de l'intelligence artificielle en matière de justice pourrait ainsi être structuré en quatre sections :

- Section Evaluation des nouveaux dossiers
- Section Accompagnement à la certification
- Section Inspection

[56] https://cs.coe.int/_layouts/15/orgchart/OrgChartCust_A.aspx?key=715&lcid=1036

- Section Recherche, développement et formation

La Section Evaluation des nouveaux dossiers serait en charge de l'instruction du dossier pour une procédure initiale de certification, du renvoi éventuel du dossier à la section Accompagnement à la certification, de la sélection des évaluateurs et des décisions et notifications d'octroi ou de refus de certification.

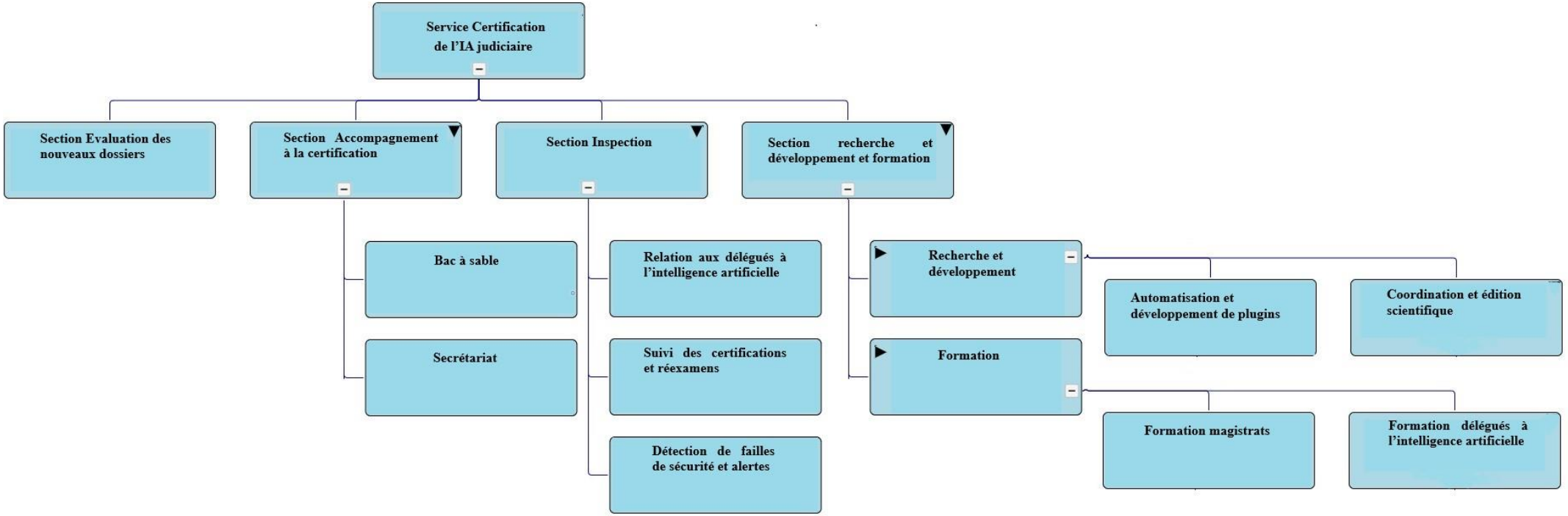
La Section Accompagnement à la certification pourrait être décomposée en deux unités l'une en charge spécifiquement du bac à sable et l'autre d'un secrétariat ayant pour vocation d'accompagner les demandeurs de certification dans la rédaction d'analyses d'impact et de les renseigner au cours de la procédure de certification. Le secrétariat serait également le point de liaison privilégié avec les éventuels organismes notifiés et les institutions partenaires.

La Section Inspection pourrait être décomposée en trois unités, la première en charge des relations avec les délégués à l'intelligence artificielle, la deuxième en charge du suivi des certifications et réexamens et la troisième en charge des failles de sécurité et alertes. Cette section serait en charge du contrôle continu et pourrait décider de la suspension ou de la suppression d'une certification.

La Section Recherche et développement et formation pourrait être décomposée en deux unités recherche et développement d'une part et formation d'autre part. L'unité Recherche et développement pourrait être subdivisée en deux parties, l'une en charge de l'automatisation et du développement de plugins pour contrôler de manière continue les intelligences artificielles et l'autre en charge de la coordination et de l'édition scientifique afin de réaliser de la veille notamment sur les techniques d'explicabilité de l'intelligence artificielle, de réaliser des études et d'envisager un rayonnement de la certification au-delà du domaine judiciaire. L'unité Formation pourrait être décomposée en deux sous-unités, la première relative à la formation et éventuellement à la certification des magistrats et la seconde relative à la formation et éventuellement à la certification des délégués à l'intelligence artificielle.

IV.2 Hypothèse d'organigramme pour un service de certification de l'intelligence artificielle judiciaire

L'organigramme présenté ci-dessous est une adaptation de celui de la Direction européenne de la qualité du médicament et soins de santé (EDQM) et de son service de certification des substances (DCEP).



V. Identification et évaluation des risques et opportunités d'une certification par la CEPEJ

V.1 Risque de concurrence et opportunités de coopération avec des projets de certifications tiers

Dans le champ spécifique du respect des droits de l'Homme par les dispositifs d'intelligence artificielle en matière judiciaire, la CEPEJ et le Conseil de l'Europe mènent un chantier unique avec la Charte et le projet de certification. En parallèle, le Comité Ad-hoc pour l'Intelligence Artificielle du Conseil de l'Europe (CAHAI) examine la faisabilité et les éléments potentiels d'un cadre juridique pour le développement, la conception et l'application de l'intelligence artificielle, fondé sur les normes du Conseil de l'Europe en matière de droits de l'homme, de démocratie et d'Etat de droit.

Dans le domaine de l'éthique de l'intelligence artificielle d'autres institutions, tout particulièrement l'UNESCO et l'OCDE (I.9), certains Etats membres du Conseil de l'Europe comme Malte et le Danemark (I.7) et des organismes de certification, comme le CEN-CENELEC ou l'IEEE (I.8) ont établi des principes convergents globalement avec ceux de la Charte et envisagent de certifier des dispositifs d'intelligence artificielle.

Le CEN-CENELEC et l'IEE pourraient être des partenaires de la certification envisagée, par exemple, pour définir des normes techniques en matière de sécurité adaptées aux enjeux spécifiques de l'intelligence artificielle dans le domaine judiciaire.

Des coopérations institutionnelles pourraient être envisagées avec d'autres institutions intergouvernementales et internationales, comme l'UNESCO et l'OCDE concernant la diffusion d'une approche de droit de l'Homme par conception pour les intelligences artificielles au-delà du champ judiciaire.

Les stratégies et premiers positionnements des Etats membres du Conseil de l'Europe, à l'image de Malte et du Danemark en matière de certification de l'intelligence artificielle pourraient permettre d'affiner les modèles de certification envisagés dans le cadre d'une consultation.

V.2 Risque d'obsolescence

Le risque d'obsolescence de la certification envisagée paraît limité dans la mesure où celle-ci est conçue pour tenir compte des avancées technologiques et de la progression spécifique de chaque intelligence artificielle certifiée. Le risque d'obsolescence est également limité du fait des typologies structurelles et pérennes (intelligences artificielles symboliques et connexionnistes) retenues pour définir les indicateurs.

Pour autant le risque d'obsolescence ne peut être écarté, par exemple, en cas d'apparition de technologies radicalement nouvelles ou avec des avancées notables dans des domaines comme l'informatique quantique, susceptibles de renforcer considérablement la puissance et les modalités de calcul pour l'apprentissage machine quantique (QML)[57] et de modifier radicalement les standards en matière de sécurité des systèmes d'information[58].

V.3 Risque de décalage par rapport aux attentes des acteurs traditionnels de la normalisation technique

La normalisation technique, de type ISO fonctionne en grande partie sur un modèle de construction de consensus entre entreprises par strates, commissions nationales (ex. AFNOR, DIN…), régionales (ex. CEN) puis internationales (ISO).

---

[57] Voir par exemple : Vedran Dunjko, Hans Briegel, « Machine learning & artificial intelligence in the quantum domain: a review of recent progress », Rep Prog Phys, 2018, 81(7):074001.

[58] Voir par exemple : 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.

Les membres des commissions de normalisation dans les domaines du numérique sont soit des dirigeants de jeunes entreprises innovantes ou de petites entreprises de conseil soit des ingénieurs détachés par des grandes entreprises ou des filiales nationales de multinationales, ce dernier cas peut être particulièrement problématique en termes de multi-représentativité pour la construction de normes (une même entreprise multinationale pouvant être représentée dans plusieurs commissions nationales). Les commissions de normalisation incluent très peu de professionnels du droit et très peu d'enseignants-chercheurs. La normalisation technique est généralement présentée comme un processus ascendant (*bottom-up*) traduisant les impératifs perçus sur le terrain par les entreprises. Les travaux des commissions de normalisation portent généralement en priorité sur les domaines porteurs économiquement et les enjeux d'interopérabilité. Dans le domaine de l'intelligence artificielle le domaine judiciaire n'est pas assez porteur économiquement (voir introduction) pour être pris en compte dans les discussions comme un cas d'usage de référence. Une certification issue de la Charte n'a donc pas de risque d'entrer en conflit avec des projets de normalisation en cours.

Les commissions nationales incluent également généralement des représentants d'un ministère, qui ont régulièrement la présidence de la commission et orientent parfois certains travaux de normalisation vers les enjeux stratégiques de leur ministère. Par exemple, en France, la commission nationale relative à l'intelligence artificielle de l'AFNOR est présidée par le Ministère de la Défense et une partie des travaux est ainsi orientée vers les armes autonomes, la gouvernance et le territoire numérique.

Au niveau européen, c'est-à-dire du CEN-CENELEC un focus group dédié à l'intelligence artificielle a été lancé en avril 2019 pour répondre aux besoins de normalisation exprimés par la Commission européenne[59]. Ce focus group s'intéresse à différents secteurs comme la manufacture, la robotique, les transports autonomes, la réalité virtuelle, la santé mais les applications dans le domaine judiciaire ne sont évoquées qu'à travers le champ relatif aux prises de décisions assistées par intelligence artificielle. Un rapprochement avec le CEN-CENELEC et la présence, par exemple, en tant qu'observateur, d'un représentant de la CEPEJ pourrait avoir pour effet d'orienter certains travaux de normalisation vers les enjeux d'une certification issue de la Charte.

Il convient également de noter que la certification envisagée issue de la Charte n'est qu'en partie technique et que cette technique est tournée vers l'éthique par conception et les droits de l'Homme dès la conception. Cette dimension paraît en partie incompatible avec une construction de consensus dominée par des entreprises commerciales. Si des coopérations sur des travaux de normalisation et de certification dans ce domaine pourraient être utilement tissées, celles-ci paraissent devoir se concentrer essentiellement sur la dimension technique. En effet, les divergences entre protection des droits fondamentaux et recherche de rentabilité commerciale paraissent des obstacles dirimants à une coopération s'entendant à la normalisation de l'éthique par conception.

V.4 Risque de qualification de barrière à l'entrée (OMC) et opportunités de réorientation éthique des pratiques

La mise en place normes techniques ou certifications difficiles à obtenir ou obligatoire (voir ci-dessus I.1), peut avoir pour effet de freiner la concurrence en rendant difficile l'accès au marché par de nouveaux acteurs, particulièrement lorsque la norme ou certification est obligatoire ou semi-obligatoire. Le risque de qualification par l'Organisation Mondiale du Commerce (ci-après « OMC ») de ces normes techniques et certifications en barrière non-tarifaire à l'entrée du marché peut également être évoqué, comme il a pu l'être dans une certaine mesure pour les Ecolabels par la doctrine, même si en raison du caractère volontaire de cette certification une plainte devant l'OMC sur ce fondement contre les Ecolabels aurait peu de chance de prospérer[60].
Au-delà du caractère volontaire de la certification, l'origine de la certification joue un rôle important dans l'analyse du risque de qualification en barrière à l'entrée et une certification issue d'une institution internationale ou intergouvernementale échappe plus facilement à ce type de qualification.
En outre, une telle certification objective, neutre et visant un objectif sociétal légitime et l'application des droits de l'Homme pourrait permettre d'éviter des stratégies et comportements d'entreprises

[59] Voir communiqué : https://www.cencenelec.eu/news/articles/Pages/AR-2019-001.aspx
[60] Kirstian Bartenstein et Sophie Lavallée, « L'écolabel est-il un outil du protectionnisme « vert » ? », Les Cahiers de droit, 2003, 44 (3), 361–393 ; Sophie Lavallée et 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 à 77.

irresponsables d'un point de vue sociétal et favoriser les solutions éthiques et conçues en tenant compte des droits et libertés fondamentaux.
Ainsi le risque de qualification de barrière à l'entrée pour une certification relative aux intelligences artificielles en matière judiciaire issue du Conseil de l'Europe paraît particulièrement faible.

V.5 Opportunités pour l'approfondissement théorico-pratique et le rayonnement de l'approche de droits de l'Homme dès la conception

La notion de droits de l'Homme dès la conception (*Human rights by design*) inscrite à l'article 1er de la Charte, dont elle est consubstantielle est porteuse de sens et paraît nettement plus opératoire que la notion répandue d'éthique dès la conception (*ethics by design*). Les droits de l'Homme dès la conception ont cette différence avec l'éthique dès la conception qu'ils s'appuient sur des dispositions, des articles auxquels le positivisme méthodologique peut donner une réelle effectivité.

Le besoin d'une prise en compte des droits de l'Homme dès la conception résulte de facteurs socio-technologiques. La méthode de droits de l'Homme dès la conception se présente comme une extension à l'ensemble des droits et libertés fondamentaux de la logique de vie privée par conception (*privacy-by-design*) qui innerve les réglementations relatives au traitement des données personnelles. L'approche par conception est rendue nécessaire par la place des dispositifs numériques et des algorithmes dans nos sociétés qui appliquent strictement des règles informatiques à la différence par exemple des magistrats qui tiennent compte au moins implicitement de l'équité. La formule de Jean Bodin « La loy sans l'équité est un corps sans âme »[61] est transposable avec une justesse renouvelée à cette situation contemporaine d'application mécanique de la règle par la machine. La méthode de droits de l'Homme dès la conception vise à injecter dans un dispositif des garde-fous contre des rigidités d'application produisant des effets contraires aux droits et libertés fondamentaux. La méthode suivie pour appliquer des droits de l'Homme dès la conception pourrait s'inspirer des méthodes suivies pour intégrer des apports jurisprudentiels à des textes de loi codifiés. Cette approche correspond à l'intégration de l'*aequitas cerebrina*, l'équité cérébrale ou non écrite, dans l'*aequitas scripta*, l'équité écrite, celle formalisée dans la loi, pour reprendre la *summa divisio* des glossateurs[62]. L'équité est traditionnellement utilisée comme un mécanisme correcteur alors que l'approche de droit de l'Homme par conception vise à supprimer autant que possible en amont le besoin de correction.

Cette méthode de droit de l'Homme par conception consiste à anticiper les dysfonctionnements susceptibles d'apparaître en pratique, à corriger certains biais, à procéder à des rééquilibrages et à formaliser de manière circonstanciée les exceptions à la règle. Les analyses d'impact axées sur la protection des droits et libertés fondamentaux participent de cette démarche de droit de l'Homme par conception, en ce qu'elles visent à anticiper les risques de mise en application.

La méthode de droits de l'Homme dès la conception si elle était appliquée avec succès à l'intelligence artificielle en matière de justice pourrait être appliquée plus largement à des intelligences artificielles utilisées dans d'autres domaines et pourrait également être potentiellement expérimentée dans les processus législatifs, après promulgation des textes de lois, pour limiter le besoin de contrôle de conventionalité devant des juridictions.

VI. Certification et responsabilités

Lors de la réunion de la CEPEJ-GT-QUAL du 18 juin 2020 les questions liées à la responsabilité relative à l'attribution de certification ont été présentées comme n'étant pas réellement spécifiques à la certification de l'intelligence artificielle en matière judiciaire.

Les enjeux de responsabilité génériques peuvent se décliner en plusieurs points :

- La responsabilisation ou déresponsabilisation des créateurs d'intelligence artificielle en matière judiciaire
- La responsabilité institutionnelle et morale de la CEPEJ et du Conseil de l'Europe en cas de certification de plateformes défaillantes
- La responsabilité pour refus d'attribution de la certification

[61] Jean Bodin, Les six livres de la République de Jean Bodin (1530-1596).
[62] Voir par exemple : « L'équité ou les équités (Journées juridiques franco-libanaises », Paris, 3-4 octobre 2002. In: Revue internationale de droit comparé. Vol. 55 N°1, Janvier-mars 2003. pp. 214-229.

Certains enjeux de responsabilité un peu moins génériques peuvent être mis en exergue comme la responsabilité du fait d'un dysfonctionnement du plugin déployé par la CEPEJ pour certifier des intelligences artificielles de manière continue ou encore la limitation de la responsabilité de la CEPEJ par le recours à un tiers certificateur pour les aspects strictement techniques, notamment ceux liés à la cybersécurité.

VI.1 Incidences et responsabilité du fait d'un dysfonctionnement d'un plugin déployé par la CEPEJ

La perspective d'un plugin déployé pour contrôler en continu les intelligences artificielles de plateformes juridiques et judiciaires permettrait un respect accru de la Charte après certification et pourrait potentiellement rayonner au-delà de ce domaine. Même si un tel plugin reposait, par exemple, sur une intelligence artificielle symbolique retranscrivant sous forme d'arbre de décisions des critères issus d'analyses d'impacts et de rapports de délégués à l'intelligence artificielle, des dysfonctionnements ou des failles de sécurité pourraient apparaître et avoir des incidences importantes sur la perception de l'utilisateur, notamment du professionnel du droit. Des décisions prises par des magistrats sur la base d'une plateforme contrôlée par le plugin, alors que ce dernier dysfonctionne et que la plateforme contrôlée n'est plus en conformité alors qu'elle apparaît comme conforme et certifiée au moment de la prise de décision pourrait réduire sensiblement la confiance dans le dispositif et la certification et créer des tensions institutionnelles.

VI.2 Limitation de la responsabilité par le recours à un tiers certificateur

Le recours à un tiers certificateur, organisme notifié, tout particulièrement pour certifier en amont, les aspects techniques notamment de sécurité informatique, en raison de l'évolution rapide de l'état de l'art et des normes en la matière, pourrait être préférable. Cette approche permettrait d'assurer la qualité de la certification et de décharger de missions accessoires l'organisme de certification éventuellement mis en place au sein du Conseil de l'Europe. Cette approche aurait également pour effet de limiter la responsabilité directe de la CEPEJ et du Conseil de l'Europe en cas de certification inadéquate dans des domaines particulièrement complexes techniquement et ne relevant pas strictement de ses missions et de la valeur ajoutée apportée en matière de certification de droits de l'Homme dès la conception.

VII. Jonctions avec la future réglementation de l'Union européenne en matière d'intelligence artificielle

L'Union européenne et tout particulièrement la Commission européenne sont très actives en matière d'intelligence artificielle que ce soit pour tenter de rendre compétitives les entreprises européennes dans ce domaine, en les finançant, par exemple avec les projets Horizon, ou pour définir les contours réglementaires de dispositifs d'intelligence artificielle conformes aux valeurs européennes.

Parmi les premières dynamiques réglementaires visant à imposer un cadre éthique à l'intelligence artificielle, la plus notable est celle du Livre blanc « Intelligence artificielle, une approche européenne axée sur l'excellence et la confiance », publié en 2020[63]. Ce Livre blanc s'inscrit dans la continuité d'une Communication de la Commission européenne intitulée « Renforcer la confiance dans l'intelligence artificielle axée sur le facteur humain »[64] de 2019 et s'articule avec les travaux du Groupe d'experts de haut niveau en intelligence artificielle (*High-Level Expert Group on Artificial Intelligence* - AI HLEG)[65].

Les travaux de la Commission européenne proposent de réglementer et d'imposer un cadre obligatoire aux intelligences artificielles à « haut risque », les autres intelligences artificielles pouvant être déployées avec une certaine liberté et se conformer à certaines règles en vue de l'obtention de certifications facultatives. La qualification d'intelligence artificielle à haut risque doit répondre à deux

[63] Livre Blanc, Intelligence artificielle Une approche européenne axée sur l'excellence et la confiance, COM(2020) 65 final.
[64] Communication de la Commission au Parlement européen, au Conseil et au Comité économique et social européen et au Comité des régions, Renforcer la confiance dans l'intelligence artificielle axée sur le facteur humain, Bruxelles, le 8.4.2019 COM(2019) 168 final.
[65] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

critères cumulatifs : un secteur à risque probable et des conséquences de l'utilisation de l'intelligence artificielle. Concernant le premier critère, l'intelligence artificielle dans le « système judiciaire » est considérée comme un domaine à haut risque dans le secteur public[66]. Concernant le second critère, le livre blanc vise explicitement « les utilisations d'applications d'IA qui produisent des effets juridiques sur les droits d'une personne physique ou d'une entreprise », ce qui est le cas, au moins pour des effets indirects, de la grande majorité des intelligences artificielles en matière judiciaire, du fait même de leur objet.

La Commission européenne prône ainsi une évaluation de la conformité objective et préalable pour garantir le respect d'exigences obligatoires pour les intelligences artificielles à haut risque[67]. Cette approche paraît parfaitement alignée avec le projet de certification envisagé par la CEPEJ mettant l'accent sur les droits de l'Homme dès la conception et la définition d'indicateurs objectifs de certification.

Les modalités d'évaluation de la conformité envisagées par la Commission européenne se réfèrent à des mécanismes existants comme le marquage CE[68], sans pour autant exclure des mécanismes nouveaux, proportionnés, non discriminatoires, transparents et objectifs[69]. Le principal inconvénient du marquage CE est qu'il est conçu pour des produits figés et non pour des dispositifs évolutifs comme les intelligences artificielles connexionnistes. Le mécanisme du marquage CE, sauf à l'adapter spécifiquement, se prête assez mal à des analyses d'impact, à des bacs à sable, au recours obligatoire à des délégués à l'intelligence artificielle ou à l'utilisation d'un plugin de certification continue en temps réel. Par ailleurs, le choix d'un recours au marquage CE aurait pour effet d'octroyer à des organismes notifiés un pouvoir d'appréciation sur des points dépassant la stricte technique et touchant aux droits fondamentaux (voir ci-dessus I.5).

Globalement l'approche de la Commission européenne pourrait converger avec un projet de certification obligatoire, plus adapté qu'un projet de certification facultative dans le domaine judiciaire[70].

A noter également que la Commission européenne envisage, par ailleurs, un soutien financier aux PME relatif aux frais de procédure et d'accompagnement pour la mise en conformité[71]. Un tel soutien pourrait être particulièrement adapté pour la mise en œuvre de bacs à sable (voir I.11).

VIII. Calendrier de déploiement et feuille de route

VIII.1 Ramification avec la « Checklist des principes de la charte dans vos traitements »

La Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ inclut une Annexe IV intitulée « Checklist d'intégration des principes de la Charte dans vos traitements ». Cette Checklist repose sur l'auto-analyse du degré d'intégration de chaque principe pris isolément dans le dispositif d'intelligence artificielle.

[66] Livre Blanc, Intelligence artificielle Une approche européenne axée sur l'excellence et la confiance, COM(2020) 65 final, p. 21, note n° 50.
[67] *Ibid.* p. 27.
[68] *Ibid.* p.28 « Les évaluations de la conformité des applications d'IA à haut risque devraient être intégrées dans les mécanismes d'évaluation de la conformité qui existent déjà pour un grand nombre de produits et de services mis sur le marché intérieur de l'UE ».
[69] *Ibid.* p.28 « S'ils n'existent pas, il peut s'avérer nécessaire d'établir des mécanismes similaires, sur la base des meilleures pratiques et des contributions éventuelles des parties prenantes et des organisations européennes de normalisation. Tout nouveau mécanisme devrait être proportionné et non discriminatoire et se fonder sur des critères transparents et objectifs en conformité avec les obligations internationales. ».
[70] *Id.*
[71] *Id.*

(*Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires de la CEPEJ, Annexe IV intitulée « Checklist d'intégration des principes de la Charte dans vos traitements », p.81*)

Il est précisé dans l'Annexe IV que cette checklist n'équivaut pas à la délivrance d'un label ou d'une certification.

Cette checklist pourrait être intégrée au stade de l'analyse d'impact par les acteurs en amont du processus de labellisation.

VIII.2 Définition des besoins initiaux de déploiement et des jalons pour la CEPEJ

Pour permettre le déploiement du projet de certification, la création d'une équipe dédiée au sein du secrétariat de la CEPEJ pourrait être envisagée.
Dans un premier temps, lors d'une phase préliminaire au déploiement d'une certification, les ressources humaines nécessaires peuvent être limitées et se concentrer sur deux objectifs d'une part le déploiement et le suivi institutionnel et d'autre part la formation et les aspects techniques.

Pour le déploiement et le suivi institutionnel un coordinateur assisté d'une équipe restreinte d'une à deux personnes, pourrait être en charge de la progression du projet de certification, de la liaison interne au sein de la CEPEJ et du Conseil de l'Europe et de la communication interinstitutionnelle et au public relative au projet de certification. Le coordinateur pourrait être chargé du suivi d'un appel à manifestation d'intérêt et de la consultation d'organismes gouvernementaux, non gouvernementaux, d'acteurs du secteur (professionnels du droit, éditeurs de plateformes et développeurs d'intelligences artificielles) et du public.

Le profil du coordinateur devrait être spécialisé dans les relations internationales et institutionnelles avec idéalement une expérience dans un organisme de certification, un organisme international axé sur le numérique ou une expérience en droit du numérique (avocat spécialiste en technologies de l'information, juriste dans une entreprise du numérique…). Les assistants, juniors ou stagiaires, pourraient avoir des profils issus de formation en communication ou relation publiques, avec idéalement une expérience préalable au sein d'organisations intergouvernementales et/ou en agence de communication numérique. Certains aspects de la communication pourraient être externalisés ou être réalisés en s'appuyant sur des ressources préexistantes de la CEPEJ.

Pour l'aspect formation et technique une équipe pourrait être composée de deux profils complémentaires d'une part un magistrat ou un enseignant-chercheur en droit du numérique, en charge de l'élaboration d'un dispositif de formation des magistrats et des délégués à l'intelligence artificielle et d'autre part un ingénieur normalisation spécialisé dans les systèmes d'information, en charge du suivi technique du projet de certification et des aspects techniques de la formation. Des universitaires et des experts techniques issus d'organismes de certification pourraient être sollicités ponctuellement pour contribuer aux travaux de l'équipe de formation et technique.

Dans un second temps, une fois le déploiement réalisé, des ressources humaines supplémentaires seront nécessaires pour chacune des sections : évaluation des nouveaux dossiers, accompagnement à la certification, inspection. L'équipe déjà créée dédiée à la formation pourrait être maintenue et éventuellement restructurée pour répondre plus spécifiquement aux besoins identifiés. Ces

recrutements supplémentaires pourraient être progressifs, en fonction du nombre de demandes de certifications, ce qui permettrait un ajustement avec les frais de dossiers demandés aux demandeurs de certifications.

Conclusion

Par l'élaboration de la Charte et la recherche d'un encadrement ou d'une certification de l'intelligence artificielle en matière de respect des droits et libertés fondamentaux, le Conseil de l'Europe, inscrit structurellement sa mission de promotion de la prise de conscience et du respect des droits de l'Homme dans le sens des évolutions techniques en s'y adaptant et en les adaptant à un projet de société.

L'utilisation de l'intelligence artificielle dans le domaine juridique et judicaire représente un enjeu sociétal important, tout particulièrement en matière de justice algorithmique. Si les avancées technologiques peuvent être porteuses d'améliorations du système judiciaire, faciliter le travail des professionnels du droit et améliorer l'accès à la justice et l'information du justiciable, une vigilance accrue est nécessaire dans ce secteur que la Commission européenne qualifie de secteur à haut risque.

La création d'une certification objective, neutre, visant l'application des droits de l'Homme dès la conception favoriserait l'émergence d'un écosystème d'intelligences artificielles conçues et déployées de manière respectueuse des droits et libertés fondamentaux. Dans un secteur à haut risque, comme le secteur judiciaire, le caractère obligatoire d'une certification apparaît comme une caractéristique faisant consensus, mais dont la mise en œuvre ne doit pas brider l'innovation.

L'initiative de certification de la CEPEJ est à la fois inédite sous l'angle sectoriel et convergente avec les travaux en matière d'éthique et d'intelligence artificielle en cours au sein d'autres organes et institutions comme l'UNESCO et l'OCDE et répond à un besoin déjà exprimé par plusieurs Etats membres du Conseil de l'Europe. Cette initiative de certification pourrait également s'appuyer sur certains points techniques des premiers travaux de normalisation réalisés, par exemple, par le CEN-CENELEC et l'IEE.

Une certification de l'intelligence artificielle en matière judiciaire permettrait également un accompagnement de projets privés et publics et l'établissement de standards à portée extra-européennes justifiant, par exemple, de faire évoluer des instruments internationaux de reconnaissance et d'*exequatur* de décisions étrangères[72] ou de sentences arbitrales[73] rendues par ou avec l'assistance d'une intelligence artificielle.

Une expérience réussie de certification de l'intelligence artificielle dans le domaine judiciaire pour lequel les considérations liées à l'éthique et aux droits et libertés fondamentaux sont essentielles pourrait utilement inspirer des certifications d'intelligence artificielle dans d'autres domaines.

[72] Par exemple, le Règlement (UE) n ° 1215/2012 du Parlement européen et du Conseil du 12 décembre 2012 concernant la compétence judiciaire, la reconnaissance et l'exécution des décisions en matière civile et commerciale.
[73] Convention pour la reconnaissance et l'exécution des sentences arbitrales étrangères, New York, 1958.

Annexes

Tableau récapitulatif des indicateurs et critères de certification				
Objectifs	**Critères**	**Modes d'évaluation**	**Cible de l'évaluation**	**Type d'IA concernée**
Traitement proportionné des données personnelles	Anonymisation des parties et intervenants personnes physiques et de leurs conseils	Consultation des jeux de données	Données non-traitées	Tous types
	Absence de notations et classements de personnes physiques ou morales sur la base de décisions de justice	Vérification de l'interface	Interface	Connexionniste
		Vérification de la base de données		
Limitation du *forum shopping*	Anonymisation du juge et de la localisation de la juridiction dans les décisions utilisées pour de la justice prédictive	Consultation des jeux de données	Données non-traitées	Connexionniste
Finalités de traitement claires	Etanchéité entre les services d'IA	Vérification des bases de données et des sources de données utilisées par chaque système	Bases de données	Tous types
Procès équitable	Mention indiquant au juge et au justiciable, le cas échéant, le caractère non explicable d'un résultat issu d'une IA (Voir également ci-dessous : Droit d'opposition du justiciable à l'utilisation d'une intelligence artificielle)	Vérification de la catégorie d'IA et vérification de la présence et de la lisibilité de la mention (Voir également ci-dessous : Vérification d'un système de notification de la décision du justiciable et de redirection effective vers une procédure classique devant un tribunal au sens de l'Article 6 de la CEDH)	Modèle d'apprentissage et interface	Connexionniste
Indépendance des juges dans leur processus de décision	Garantie contre le profilage des juges	Vérification A/B testing des résultats de recherche	Moteur de recherche et données traitées	Tous types
	Adéquation entre le critère affiché et le mode de classement effectif des résultats de recherche	Vérification par audit des résultats de recherche	Moteur de recherche	Tous types
	Transparence des pondérations entre critères en cas de recherches multicritères	Vérification de l'existence de mentions explicatives	Interface et moteur de recherche	Tous types
		Audit des résultats de recherche		
	Transparence des critères utilisés en cas de recherche dite par « pertinence »	Vérification de l'existence de mentions explicatives	Interface et moteur de recherche	Tous types
		Audit des résultats de recherche		
Ethique et droits de l'Homme dès la conception	Absence d'atteinte aux droits de l'Homme	Rapport de présentation des arbres de décision expliquant la prise en compte des droits et libertés fondamentaux	Rapport	Symbolique

Tableau récapitulatif des indicateurs et critères de certification				
Objectifs	**Critères**	**Modes d'évaluation**	**Cible de l'évaluation**	**Type d'IA concernée**
	Absence d'atteinte aux droits de l'Homme	Rapport de présentation des données d'entraînement et des modalités d'entraînement expliquant la prise en compte des droits et libertés fondamentaux	Rapport	Connexionniste
Non-discrimination fondée sur des données sensibles	Suppression des marqueurs rattachables aux données sensibles des parties (domicile, revenu, situation familiale, capital social)	Vérification par consultation des jeux de données	Données non-traitées	Non contrôlée par un organisme public
		A/B testing à partir d'information et marqueurs rattachables à des données sensibles en changeant, le cas échéant, lors de chaque test un des paramètres suivants : nom, domicile, revenu, situation familial, capital social, élément de contexte spécifique pertinent	Moteur de recherche	
		Formulaire donnant la possibilité pour les utilisateurs d'exprimer de manière circonstanciée des demandes de suppression de marqueurs rattachables à des données sensibles avec copie au délégué à l'intelligence artificielle, le cas échéant, et à l'autorité de contrôle (entité attribuant la labélisation).	Interface	
	Dispositifs garantissant le caractère nécessaire et proportionné du traitement	Analyse d'impact par l'organisme de traitement expliquant les garanties mises en œuvre	Rapport	Contrôlée par un organisme public
		Désignation d'un délégué à l'intelligence artificielle	Lettre de nomination	
Qualité et sécurité de la donnée	Obtention préalable de la Norme ISO/IEC 27001	Vérification de la certification norme ISO/IEC 27001	Attestation d'un organisme notifié	Toutes
Transparence	Code open source ou transmission du code source sous couvert de confidentialité	Vérification de l'entièreté du code source	Code source	Toutes
Garantie à l'utilisateur la maîtrise du dispositif	Assistance téléphonique disponible et pertinente ; interface ergonomique et intuitive ; des guides d'utilisation clairs et détaillés, des tutoriels courts axés sur la prise en main de chaque fonctionnalité ; des formations à l'outil	Vérification de la présence sur l'interface	Interface et test de fonctionnalité	Toutes

Tableau récapitulatif des indicateurs et critères de certification				
Objectifs	**Critères**	**Modes d'évaluation**	**Cible de l'évaluation**	**Type d'IA concernée**
Maîtrise par l'utilisateur et procès équitable	Information lisible pour le justiciable de l'utilisation d'une intelligence artificielle pour son affaire	Vérification de la présence d'une bannière d'information mentionnant l'utilisation d'une intelligence artificielle devant être cliquée avant tout accès au service	Interface	Toutes
	Droit d'opposition du justiciable à l'utilisation d'une intelligence artificielle	Vérification de la présence d'un module d'expression du consentement à la dernière ligne du déroulé de la bannière d'information	Interface	Toutes
		Vérification d'un système de notification de la décision du justiciable et de redirection effective vers une procédure classique devant un tribunal au sens de l'Article 6 de la CEDH	Interface et test de fonctionnalité	

Outils complémentaires	
Désignation	**Objectif**
Plugin d'intelligence artificielle interprétable de contrôle des intelligences artificielles dans le domaine judiciaire	Contrôle continu des IA connexionnistes et appui aux délégués à l'IA
Registre blockchain pour l'immatriculation et la signature des intelligences artificielles certifiées	Création de traces permettant l'identification d'IA
Registre blockchain des décisions de justice	Intégrité des données et respect des exigences d'anonymisation
Bac à sable	Accompagnement des demandeurs de la certification lors de la procédure
Formation pratique des magistrats à l'intelligence artificielle en matière judiciaire	Sensibilisation aux enjeux de la Charte et de la certification, acquisition de compétences techniques et partage de bonnes pratiques