Le Ius artificiale entre intériorité et boîte noire : Le droit de l'IA est-il soluble dans le droit ?

Géraldine Aïdan, Primavera de Filippi

LE *IUS ARTIFICIALE* ENTRE INTÉRIORITÉ ET BOÎTE NOIRE : LE DROIT DE L'IA EST-IL SOLUBLE DANS LE DROIT ?

Geraldine Aïdan (geraldineaidan@hotmail.fr)
Primavera De Filippi (pdefilippi@gmail.com)

Centre d'études et de recherche en sciences administrative et politiques (CERSA)
CNRS — Université Paris II

Cet article fait partie d'une recherche plus globale visant à démontrer que nous assistons aujourd'hui à l'émergence d'un nouveau droit—le ius artificiale[1]—issu de l'adoption de l'Intelligence Artificielle (IA) dans différents domaines du droit : outil de prédiction pour la doctrine et les professionnels du droit[2], outil d'aide à la décision judiciaire et outil de décision juridique. L'étude présentée ici a pour objectif de questionner la compatibilité entre le Ius artificiale et le droit humain, en nous intéressant tout spécialement à l'intériorité du juge et en interrogeant sa possible transposition au sein d'un système d'IA. En mettant en lumière l'existence d'une boîte noire mobilisant une interprétation spécifique du droit et du monde par les systèmes IA, nous montrerons que le Ius artificiale est doté de spécificités qui le rendent potentiellement incohérent et incompatible avec le droit humain.

Les applications de l'IA dans le droit datent d'une vingtaine d'années mais se généralisent désormais en Europe et dans les pays anglophones, aussi bien dans les systèmes de common law que de civil law.[3] Ainsi, l'IA tend à devenir un auteur à part entière de normes juridiques qui sont susceptibles d'être intégrées au sein du droit positif. On distingue aujourd'hui notamment deux systèmes d'IA appliqués au droit : les systèmes experts (SE) et les systèmes d'apprentissage automatique (machine learning ou ML).

Les systèmes experts sont des systèmes de raisonnement fondés sur des arbres de décision construits à partir de connaissances spécialisées. Ils s'appuient sur des bases de connaissances pré-existantes, développées par des experts, afin de tirer des conclusions sur une question donnée à partir de certains

---

[1] "Le *Ius Artificiale, quel nouveau droit pour demain ?"* ouvrage en cours d'écriture par les même auteurs

[2] Voir "La Justice prédictive", *Archives de philosophie du droi*t, numéro 60

[3] Il existe de nombreuses applications de l'IA dans différents domaines du droit : des outils pour faciliter la recherche de documents juridiques et de jurisprudence (e.g. Westlaw Answers, Lexis Advance, ROSS); des applications pour optimiser l'audit de contrats (e.g. LawGeex AI, Kira Systems, Luminance, Hyperlex) ; des outils d'aide à la décision et des applications de justice prédictive (e.g. COMPAS, Predictice, Legalmetrics); ou encore des systèmes de support à l'arbitrage international (e.g. ArbiLex, Jus Mundi).

paramètres d'entrée. Ces systèmes utilisent un langage symbolique et des mécanismes de raisonnement qui simulent le raisonnement des humains. Alors que les systèmes experts ne posent pas spécialement de problèmes en termes de transparence et d'intelligibilité, ils sont cependant limités dans la mesure où les règles qui les régissent doivent être prédéfinies à l'avance par des experts, et ne peuvent donc pas évoluer au fur et à mesure que de nouvelles informations sont fournies au système.

Les systèmes d'apprentissage automatique sont, eux, des outils d'analyse statistique qui s'appuient sur des réseaux de neurones artificiels (RNA) afin de produire des connaissances inductives par rapport à un ensemble de données issues et dérivées des décisions de justice et administratives. Il existe des systèmes d'apprentissage automatique supervisés et non supervisés : les premiers se fondent sur des éléments d'apprentissage qui leur sont fournis par des humains, les seconds vont déterminer par eux-mêmes quels sont les éléments déterminants de chaque jeu de données par rapport à une décision. L'avantage de ces systèmes, par rapport aux systèmes experts, est qu'ils sont capables d'identifier de nouveaux critères ou des corrélations auxquels les humains n'avaient pas pensé. Par contre, ils posent des problématiques d'intelligibilité car il est difficile (voire impossible) pour un humain d'accéder à la boîte noire de l'IA afin de comprendre le choix de la décision rendue.

On s'interroge notamment sur la capacité de ces systèmes d'intelligence artificielle à appréhender (et donc à représenter) la fiction juridique d'un juge humain au sein d'un système informatique—ce qui implique une "codification" de deux fictions: celle du monde réel, et celle du droit. Nous entendrons par "fiction juridique" la construction spécifique d'une réalité par un organe du droit, qui s'écarte de celle admise par un système de pensée de référence.[4] Ainsi, nous nous rapprochons de la conception de la fiction comme « imaginaire social instituant"[5] sans pour autant nous y rattacher. En effet, si nous admettons que le droit "est constitué aussi de narrativité, c'est-à-dire de grands et de petits récits, qu'il est tissé de fictions »[6], l'angle d'approche ici adopté est celui du normativisme juridique.[7] Nous nous concentrons sur les modes d'élaboration de la norme juridique en tant qu'elle mobilise des fictions différentes selon la nature de l'auteur de celle-ci: l'humain ou l'IA.

---

[4] Par exemple, un juge va appréhender le concept d'"arbre" différemment d'un botaniste. Nous élargissons ainsi de la définition classique de la fiction en droit entendu comme étant un procédé juridique consistant " à prendre le faux pour le vrai" c'est à dire à " d'abord à travestir les faits, à les déclarer autres qu'ils ne sont vraiment, et à tirer de cette adultération même et de cette fausse supposition les conséquences de droit qui s'attacheraient à la vérité que l'on feint, si celle ci existait sous les dehors qu'on lui prête" Yann Thomas, les opération du droit, EHESS Seuil, Gallimard, 2011, p. 133.

[5] Fr. Ost, Raconter la loi. Aux sources de l'imaginaire juridique, op. cit., p. 34. Voir pour une approche critique voir O. Pfersmann, "les modes de la fiction : droit et littérature" in *Usages et théories de la fiction* (coord. F. Lavocat), PUR, p . 39-61

[6] J. Michel, « François Ost, Raconter la loi (aux sources de l'imaginaire juridique) »François Ost, Raconter la loi (aux sources de l'imaginaire juridique) », *RTDciv*., 2005, p. 499,

[7] Le normativisme a pour ambition de décrire le droit sans considérations morales ni politiques. Son objet se compose de normes juridiques, étant entendu que la norme est la « signification d'un acte » (H. Kelsen, Théorie pure du droit, op. cit., p. 13) ou plutôt la signification« d'énoncé déontiquement modalisé », (O. Pfermsmann, « Pour une typologie modale de classe de validité normative » op. cit., p. 74) par lequel un comportement humain est obligatoire, permis ou interdit et qu'une norme est juridique lorsqu'une autre norme juridique lui attribue cette qualité (Cf. KELSEN (Hans), Théorie pure du droit, op. cit., p. 16). Au-delà des différentes conceptions assignées au droit, nous entendrons celui-ci comme l'ensemble des normes juridiques.

On s’interroge ici sur la capacité des systèmes d’intelligence artificielle à appréhender—et donc à représenter—les fictions du juge humain au sein d’un système informatique. Bien que les systèmes experts et les systèmes de machine learning, ces mobilisent des stratégies différentes,nous revendiquons que ces derniers sont tous les deux incapables de transposer correctement ces fictions dans leur propre système de fonctionnement.

D’une part, en ce qui concerne la fiction du réel, les systèmes d’IA ne peuvent pas appréhender le monde comme les humains, car ils ne sont pas dotés d’un corps qui aide à l’élaboration d’une vie psychique intérieure—ce que l’on appelle l"intériorité"[8]. Ils disposent cependant de mécanismes leur permettant d’analyser et de traiter les informations qu’ils reçoivent afin de construire leur propre représentation du réel dans un système informatif—un système qui comportera une fiction nécessairement différente de celle des humains. D’autre part, pour ce qui en est de la fiction du droit, nous allons démontrer que l’objectif commun des systèmes d’IA n’est pas de “répliquer” le fonctionnement d’un juge humain, mais uniquement de le “simuler” : de faire “comme si”[9] c’était un juge humain qui jugeait. Pour y arriver, les systèmes experts et les systèmes de machine learning mobilisent des moyens différents: les premiers tentent d’expliciter “l’intériorité juridique” du juge humain dans un système informatique, sans pour autant fabriquer une nouvelle boîte noire (I), les seconds construisent une nouvelle boite noire artificielle, sans être capable d’élaborer une fiction partageable avec les humains (II). C’est ce dernier constat qui nous interpelle. Nous partons en effet de l’hypothèse que la fonction même de juger implique l’existence d’une intériorité psychique impliquant une représentation du monde et du droit suffisamment partagée entre le producteur de normes juridiques et ses destinataires.

I. Systèmes Experts: une fiction sans boîte noire

Comment le système SE se représente-t-il le droit quand il juge ? Quel est le contenu de son “intériorité” juridique ? Le concepteur du SE, aidé par des juristes (“les juristes-concepteurs”), tente d’expliciter le raisonnement de ce que l’on peut appeler un “juge-type” dans un système informatique clair et facilement interprétable, mais nécessairement incomplet. Par conséquent, il ne constitue qu’une représentation partielle et altérée, voire dénaturée de la fiction du droit par ce juge-type. Cette représentation du droit s’incarne dans une codification du droit qui se réalise en deux étapes : l’explication du raisonnement d’un juge-type (1) et sa traduction dans un système informatique (2) .

1. L'émergence du juge-type

[8] Le concept d’intériorité sera défini de manière provisoire comme visant l’ensemble des phénomènes attribués intuitivement comme étant « à l’intérieur » d’une entité et constitutifs d’une forme de subjectivité, de manière réelle, simulée, supposée ou posée par convention.

[9] Hans Vaihinger,La philosophie du comme si, préface et première traduction française de Christophe Bouriau, Cahier spécial 8 de la revue Philosophia Scientiae éd. Kimé Paris 2008, 339

Le SE a comme point de départ l'intériorité d'un juge-type, et tente de l'interpréter et de la codifier dans un langage informatique. Ainsi, la conception d'un SE implique la nécessité d'expliciter, dans un système informatique strict et formel, les concepts du droit et du réel cachés dans "*l'intériorité juridique*" des juges humains, qui est par définition inaccessible. Les juristes concepteurs vont tenter de formaliser au sein du SE la manière dont un juge-type se représente l'ensemble du droit (comprenant à la fois les normes générales et abstraites, et les normes individuelles et concrètes). Or, bien que le but du SE soit de répliquer de la manière la plus fidèle possible le raisonnement de ce juge-type, il ne peut le faire que de manière nécessairement incomplète.

Techniquement, cette première étape consiste en une mise en abstraction des règles juridiques dans une structure de pensée plus formelle, par exemple par la constitution d'un arbre de décision. Cet arbre de décision implique une reconfiguration du système juridique en un système formel qui traduit (et fait d'ailleurs ressortir) les contours et les caractéristiques de la fiction du droit que le juriste-concepteur entend articuler dans le système expert. Il s'agit en quelque sorte de modéliser une certaine représentation mentale du droit, en déclinant les règles juridiques jusqu'à leur éléments les plus infimes. Selon les différentes déclinaisons possibles, un même système juridique pourra alors donner lieu à de multiples fictions (multiples interprétations du droit par le juge), et chacune de ces fictions donnera alors lieu à une abstraction et une modélisation correspondante.

Pour ce faire, le juriste-concepteur doit d'abord identifier des normes pertinentes dans le domaine d'étude. La construction d'un SE présuppose en effet le besoin de répondre à une question donnée (par exemple, quelles sont les normes pertinentes à la détermination du statut juridique d'une œuvre de l'esprit ?). La réponse à cette question se construit à partir d'un arbre de décision qui modélise une fiction particulière du droit par rapport à cette question à traiter.

Cette identification des normes pertinentes est suivie ensuite d'une hiérarchisation générale et abstraite des normes juridiques. Ce procédé est nécessairement effectué ex-ante (i.e. avant et indépendamment de l'appréciation de ces normes à l'occasion d'un cas d'espèce). Ainsi, le juge concepteur sera obligé d' expliciter et de figer plusieurs significations d'un énoncé normatif en les déclinant abstraitement dans une multitude de cas d'espèces potentiels.

2. Traduction dans un système informatique

La deuxième étape consiste en une instanciation de cette première abstraction de la fiction du droit par le juge élaboré lors de la première étape. Nous entendons par instanciation la manière dont les développeurs du système vont codifier la représentation abstraite des juristes-concepteurs (c'est-à-dire . l'arbre de décision qui a été élaboré préalablement) en la formalisant au sein d'un langage informatique strict et formel. La traduction de la représentation humaine du réel et du droit dans un langage informatique implique alors une re-fictionnalisation[10] du réel et du droit, par le biais d'une

codification informatique qui peut être interprétée par une machine.

Cette tentative de codification des concepts humains (afférent à des phénomènes juridiques ou réels) entraîne une "atomisation", voire une fragmentation de ces concepts en une série de paramètres en entrées indépendants les uns des autres (et qui sont donc incapables de refléter ces concepts de manière unitaire). Le SE fonctionne alors sans concepts: contrairement au ML, il ne prend en compte qu'une multitude d'éléments fragmentés et codifiés de manière indépendante.

3. Implications

Nous analysons ici les implications du modèle SE sur sa capacité à simuler la fonction de juge. Celle-ci apparaît limitée en raison de son incomplétude (a), de la nécessaire résolution ex-ante de tout conflit de normes (b), ce qui implique une interprétation figée du droit, et donc une perte de pouvoir discrétionnaire du juge (c).

a. Incomplétude

Comme nous l'avons vu auparavant, le système expert tente d'expliciter l'intériorité d'un juge-type (qui incarne une représentation particulière du droit telle qu'elle est perçue par les juristes concepteurs) de manière abstraite et formelle, en un langage informatique. Or, cette représentation de la fiction du droit par le juge apparaît nécessairement incomplète (et potentiellement erronée) pour deux raisons principales.

D'une part, techniquement, cette représentation du droit dans un langage informatique ne peut être entièrement fidèle à la fiction élaborée par le juge humain car elle est corsetée dans une reformulation dans des clauses de type "si A, alors B", qui exigent une détermination préalable et une déclinaison précise des paramètres en entrée (A) et des conséquences en sortie (B). Ainsi, tout concept juridique doit être codifié en listant explicitement l'ensemble des éléments visés par celui-ci. Par exemple, la traduction d'un concept juridique compris dans un énoncé normatif (ex. une œuvre littéraire) se fera par l'établissement d'une liste d'éléments singuliers et distincts (par ex. "une monographie", "un article scientifique", "un poème" mais pas "une recette")—et donc nécessairement limitée ou incomplète. Seuls ces éléments pourront alors être traités en tant que paramètres par le SE. D'autre part, en plus de cette traduction de concepts juridiques dans une série de paramètres, la transposition de la fiction d'un juge-type dans un SE implique une reformulation de normes juridiques en un langage strict et formel (avec des clauses de type "si A, alors B"), dont l'interprétation est forcément limitée par rapport aux clauses écrites en un langage naturel. Le langage informatique ne peut traduire l'ensemble des nuances et des ambiguïtés du langage naturel; il y a donc une part d'indétermination propre au langage naturel que le SE ne peut appréhender.[11]

---

[10] La question sera alors de savoir si ce processus de fictionnalisation dans le SE va donner lieu à une métafiction (qui est une simple transposition—bien que partiellement dénaturée—de la fiction du juge type) ou à une nouvelle fiction indépendante de la fiction du juge-type humain

b. Résolution ex-ante de tout conflit de normes

La formalisation du raisonnement d'un juge-type au sein d'un SE comporte aussi une hiérarchisation des normes dans un arbre de décision qui régit le système. Cette hiérarchisation ne porte plus seulement sur les normes appartenant à différents rangs normatifs (la constitution est supérieure à la loi, qui est supérieure au règlements, qui sont eux supérieurs à la jurisprudence, etc.) mais aussi sur l'ensemble des normes appartenant à un même rang normatif. Dans ce dernier cas, le juriste concepteur devra imaginer tous les conflits de normes potentiels qui peuvent se révéler dans le traitement d'une question juridique donnée, et les résoudre à l'avance (ex-ante), en déclinant chaque norme juridique en une série de normes individuelles et concrètes structurées de façon à éviter tout possible conflit.[12] De multiples autres représentations (concurrentes) de la fiction du droit par un juge-type sont alors envisageables, en fonction de la manière dont les juristes concepteurs vont décider d'interpréter les normes juridiques et de concevoir puis résoudre tous les conflits de normes qui pourraient se présenter.

c. Fixation et perte du pouvoir discrétionnaire du juge

La représentation sélectionnée par les concepteurs du SE est alors figée au sein d'un arbre de décision qui va déterminer la façon dont le SE gérera certaines questions juridiques, sans possibilité de réinterpréter les normes ou de les hiérarchiser à la lumière du cas d'espèce en question. Une fois fixé, il ne sera donc plus possible de réordonner les normes juridiques ainsi codifiées, ou d'appliquer les normes juridiques aux cas d'espèces qui n'ont pas été envisagés préalablement par le juriste concepteur, sous réserve d'une nouvelle intervention de celui -ci.

Tel qu'il est mis en place dans un système expert, le *Ius artificiale* limite alors la place de l'humain : la décision humaine est prévue *ex-ante* (a priori) de façon abstraite et générale, au lieu d'être élaborée *ex-post* (a posteriori) par rapport à chaque cas concret. Le SE retire ainsi une partie du pouvoir discrétionnaire, traditionnellement assigné aux juges.

II. Machine Learning: une boîte noire sans fiction

---

[11] En cela, il n'y a pas d'interprétation possible dans le SE, au sens où la classique interprétation consiste en la détermination de la part d'indétermination d'un énoncé normatif : ici la part d'indétermination est muselée: il y n'a que du déterminé.

[12] Exemple : On ne peut limiter la liberté d'expression, dans tous les cas où situation (a), (b), (c) versus droit à la vie privée;

Comment les systèmes d'apprentissage automatique (machine learning) se représentent-t-ils à leur tour le droit quand ils "jugent", et quel est le contenu de leur intériorité juridique ? Alors que le juriste concepteur du SE est responsable d'identifier et de formaliser au sein d'un système informatique la fiction du réel et du droit d'un juge-type, le concepteur d'un système ML se limite à sélectionner les données qui seront utilisées pour entraîner le système, et à choisir le logiciel de ML à utiliser pour traiter ces données. Le système apprendra ensuite "par lui-même" à partir de ces données afin de dériver ou d'inférer (à partir de corrélations) des règles qui sembleraient gouverner l'ensemble des cas présentés au système. Le ML n'a donc pas vocation à simuler un "juge-type" dans sa fonction de juger (i.e. quel serait le jugement d'un juge "idéal"?); mais plutôt à simuler un "juge-moyen" appliquant les règles ainsi identifiées (i.e. quel serait le jugement statistiquement attendu par rapport à tous les juges ayant pris une décision sur un cas d'espèce similaire?).

Le ML va donc inférer un nouveau système de règles entièrement déconnectés du système juridique sur lequel il est censé intervenir, car appartenant à une nouvelle fiction du droit, générée par l'IA. Cette nouvelle fiction du droit est créée à partir d'un double processus: (1) l'émergence d'un juge-moyen par l'analyse statistique des paramètres en entrée afin d'identifier les règles "moyennes" (statistiquement signifiantes) qui vont devenir une partie intégrante de la fiction de l'IA, et (2) la création d'une nouvelle boîte noire qui incarne cette fiction artificielle de l'IA dans un réseau de neurones inaccessible à l'humain.

1) L'émergence du "juge-moyen" par l'élaboration d'une "loi sociale du droit"

Le ML n'effectue pas une traduction des règles de droit en un langage informatique, mais effectue plutôt un travail de "sociologie du droit" qui a vocation à faire émerger une "loi sociale"- parallèle à la norme applicable ou appliquée par les juges humains - à partir de laquelle il tente de simuler l'activité d'un juge-moyen.

Afin de réaliser cette "sociologie du droit" le ML ne se nourrit pas des règles de droit positif telles qu'elles sont détaillées dans les lois, les normes de l'administration, la jurisprudence de principe (comme le ferait un juge humain), mais uniquement de la jurisprudence. Ainsi, le ML tente de "simuler" l'activité d'un juge humain-moyen en fabriquant des corrélations entre les paramètres en entrée (les faits du cas d'espèce en question) et les décisions juridiques associées à des cas similaires. Son objectif est de faire ressortir la fiction "moyenne" (i.e. celle partagée par l'ensemble des juges d'une communauté donnée à un moment donné).

Ainsi, à la différence du SE, qui tente d'expliciter la fiction du réel et du droit par le juge avec la création d'un système représentant le raisonnement d'un juge-type, le ML va recréer une nouvelle boîte noire (indépendante de la boite noire du juge) et fabriquer une nouvelle fiction (du réel et du droit) qui appartiendrait à un "juge-moyen". Le ML ne tente pas d'expliciter la fiction du juge, il se limite à vouloir "simuler" la décision qu'un juge-moyen proposerait selon une procédure statistique (en observant les relations et corrélations entre les paramètres en entrée et les décisions juridiques). Le ML va donc créer sa propre fiction du réel et du droit (indépendante de la fiction du juge) dont le but n'est pas de codifier le droit, mais uniquement de prendre des décisions "comme si" il s'agissait d'un juge (statistiquement) moyen.

Cette fiction ainsi induite par le ML (suite à l'élaboration d'une loi sociale) permet de faire ressortir les biais des juges dont les décisions ont été traitées par le système. Comme outil descriptif, le ML apparaît alors comme un allié précieux du juge humain et des acteurs du droit pour identifier ces biais humains. Cependant, en tant qu'outil normatif (producteur ou co-producteur d'une norme), le ML mobilise des règles générales nécessairement erronées: non seulement car ne sont pas celles du droit positif de référence mais aussi car elles relèvent de la loi comme système "social" et non comme "système idéal". Le ML se fonde en somme sur un système de normes parallèles (non juridiques) induites par lui-même suite à une analyse statique des décisions de jurisprudence qui lui ont été fournies.

Ces règles sont par ailleurs déjà "faussées" en tant que telles car elles sont induites à partir d'un cercle des normes extrêmement réduites (la jurisprudence) et non représentative du droit dans son ensemble. Elles se fondent en effet principalement sur des règles qui ne sont pas "effectives" (i.e. qui ne sont pas correctement appliquées par la société ou qui posent problème dans son application) et qui donnent alors lieu à un contentieux. Les règles effectives sont sous-représentées dans le corpus à partir duquel le système ML va construire sa propre conceptualisation du droit.

2) Création d'une nouvelle boîte noire : une fiction artificielle inaccessible aux humains

En essayant de simuler les activités d'un juge moyen, le système ML construit sa propre fiction artificielle du droit, une nouvelle boîte noire "artificielle" avec ses propres concepts normatifs du réel qui n'ont (presque) rien à voir avec ceux d'un juge humain. Le ML apprend, en effet, à partir des faits et des décisions de justice, et construit sur cette base une nouvelle boîte noire qui lui permettra de "simuler" les activités du juge. Or, bien que l'objectif soit de produire ou proposer en définitive une norme qui "ressemble" suffisamment à celle qu'aurait prise un juge-moyen, il n'y a, en réalité, aucune

tentative de traduire ou d'expliciter l'intériorité d'un juge. Il en résulte deux fictions (celle du juge humain et celle du ML) qui coexistent "comme si" l'une était le reflet de l'autre[13], alors qu'elles sont relativement indépendantes et potentiellement incompatibles. D'une part, le ML se base pour sa construction sur la jurisprudence uniquement (et non pas les normes générales et abstraites) duquel il fait émerger une "loi sociale" du droit. D'autre part, le traitement qu'il fait de la jurisprudence est un traitement nécessairement incomplet, qui ne prend en compte que les informations—codifiées sous formes de données compréhensibles par une machine—que le juriste-concepteur lui a fournies.

Dans un premier temps (durant la phase d'apprentissage) le système va générer des modèles et des corrélations à partir d'une série de paramètres en entrée qui ont été préalablement annotés par des experts.[14]. L'apprentissage se matérialise au sein d'un réseau de neurones artificiels (RNA) par l'évolution des valeurs constituant sa structure interne (i.e. le poids des connexions entre les neurones). Ce processus implique la création d'une nouvelle boîte noire utilisant des règles potentiellement dé-corrélées des normes juridiques en vigueur.

Dans un deuxième temps (durant la phase d'application), le système va utiliser son propre réseau de neurones afin de reconnaître certains "patterns" au sein de nouvelles situations, pour en dériver une décision de justice (e.g. pour décider si les faits d'un nouveau cas comportent un tort).[15] Or, puisqu'il s'agit d'une nouvelle boîte noire, on ne peut évaluer que le résultat de ce processus de raisonnement, car un système ML n'est pas en capacité de fournir de motivation pour expliquer comment il est arrivé à ce résultat. En effet, alors que les systèmes experts fournissent un degré élevé de transparence et d'intelligibilité, dans le cas des systèmes d'apprentissage automatique, il est beaucoup plus compliqué d'identifier le raisonnement qui a conduit à une décision donnée, et de connaître les motivations sous-jacentes à cette décision. Ces systèmes utilisent un langage dit sub-symbolique, où les concepts utilisés dans le réseau de neurones sont exprimés sous forme de vecteurs ou de formules mathématiques qui—même s'ils peuvent être identifiés en tant que tels—ne sont pas traduisibles en un langage compréhensible par les humains. L'intelligibilité des mécanismes d'apprentissage automatique est ainsi fondamentalement limitée à cause de ces différents registres de langage : si l'on peut tracer le raisonnement global de ces algorithmes, les paramètres et les corrélations créés au sein de leurs réseaux de neurones ne seront pas pour autant directement compréhensible pour les humains. Au fur et à mesure que ces systèmes enrichissent leur propre langage avec de nouveaux paramètres et

[13] La question sera alors de savoir si la simulation porte exclusivement sur la solution proposée (i.e. simuler la décision d'un juge-moyen) ou sur tout le processus normatif y compris la mise en fiction (i.e. simuler le raisonnement d'un juge-moyen).

[14] Par exemple, dans le cas d'une décision juridique, les paramètres en entrée seront annotés avec les éléments juridiques - des faits ou des normes - pertinents à la décision.

[15] Rapport de la Commission de réflexion sur l'Éthique de la Recherche (20017) Éthique de la recherche en apprentissage machine, Juin 2017.
http://cerna-ethics-allistene.org/digitalAssets/52/52472_CERNA_Ethique_de_la_recherche_en_apprentissage_machine.pdf

corrélations, se crée alors un décalage croissant entre les modalités de raisonnement humain et les mécanismes de raisonnement algorithmique. Ainsi, bien que les ML nous permettent d'identifier de nouveaux critères ou corrélations auxquels les humains n'avaient pas pensé, il construisent nécessairement leur décision à partir d'une boîte noire totalement étrangère à l'intériorité que mobilise le juge humain (c.f. inaccessibilité du raisonnement et des motivations sous-jacentes) . Cette nouvelle boîte noire repose sur une fiction qui apparaît alors comme *a-humaine*— elle est construite sans l'humain et, une fois créée, elle reste inaccessible aux humains.

3) implications

a. Confusion entre "être" (faits) et "devoir être" (droit)

Le "raisonnement" ou le mode de fonctionnement du ML se fonde sur une confusion ou identité entre droit et réalité (*devoir être* et *être*). Le ML ne reçoit que des paramètres en entrée sous formes de "faits" et "décisions" traités indistinctement comme des informations factuelles. Les normes générales du droit ne lui sont jamais communiquées. Le ML construit ainsi une autre mise en cohérence des éléments du réel et du droit humain: il crée son propre système de "normes" à partir de ce corpus qui ne se situe qu'au niveau factuel, en fabriquant une "loi sociale du droit".[16] Ce système de normes parallèle au droit ne peut être inféré qu'à partir de ce qui a déjà existé juridiquement (les affaires jugées, les faits retenus par le droit). Le ML ne peut pas imaginer d'autres faits que ceux déjà jugés dans une décision. En définitive, il adopte une approche exclusivement descriptive (*être*) plutôt qu'une approche prescriptive (*devoir être*).

Primo, le ML saisit un ensemble de données composé d'une part, des faits sélectionnés par le concepteur (e.g. caractéristiques physiques, nationalité, et comportements de l'accusé, informations contingentes, etc.) ainsi que les décisions jurisprudentielles ou administratives associées à ces faits (e.g. coupable / non coupable, responsable / non responsable...), à partir duquel le ML va procéder à la construction de corrélations, qui ne sont pas soutenues par une logique d'imputation (*devoir être*) mais uniquement par une logique de probabilité (*pourrait être*).

Secondo, le ML va agréger ces corrélations afin d'en tirer une loi sociale à partir de laquelle il va traiter de nouveaux cas d'espèces. Or, cette "loi" ne relève pas du *devoir être*, elle ne se situe pas au niveau prescriptif (i.e. elle ne dit pas ce qui doit être), ni au niveau descriptif (i.e. elle ne dit pas ce qui *est*), mais plutôt au niveau probabiliste (elle dit ce qui "*devrait être*" au regard de ce qui "*pourrait*

[16] Nous entendons par "loi sociale du droit", la loi implicite, sous-jacente qui commande statistiquement la résolution des cas d'espèces. Il s'agit de la loi qui émane de l'analyse statistique des décisions traitées dans un domaine donné.

être"). Cependant, le devoir être contenu dans la "décision" de l'IA ne relève plus d'une norme idéale (tel comportement doit être selon un système de valeurs) mais d'une norme sociale (tel comportement devrait être statistiquement selon ce qui a été).

b. Raisonnement par corrélation (et non pas causalité)

Le ML n'applique par le droit humain. Il va induire à partir des éléments du droit et du réel que le juriste-concepteur lui a donné une nouvelle loi parallèle (et inaccessible à l'humain) qui va mettre en corrélation ces éléments du réel et du droit. Ainsi, bien que le ML ait vocation à simuler un juge-moyen, il s'appuie pour ce faire sur un système normatif qui lui est propre (induit par une sociologie du droit) qui est nécessairement différent du droit positif. Puisque le ML n'a pas accès aux liens d'imputation exprimés par les normes juridiques, il ne peut se focaliser que sur des liens de corrélations exprimées par les informations factuelles qui lui sont données.

De plus, étant donné l'incapacité du ML à appréhender les liens de causalité, ce processus de construction normative résulte d'une procédure inductive et non pas déductive et entraîne des conséquences sur la nature de la fiction en jeu. Alors que le juge humain crée sa fiction à partir d'une interprétation déductive du droit, à la lumière de sa conceptualisation du monde - partagée par les autres juges - le ML créé sa fiction[17] à partir d'une interprétation inductive du droit, à la lumière d'une conceptualisation extrêmement limitée du réel. Il n'y a alors pas en tant que telle de représentation du monde et du droit, mais une mise en cohérence des éléments reçus et captés. Par conséquent, la fiction apparaît totale dans les systèmes ML : une fiction parallèle et *a-humaine*, qui existe indépendamment de la réalité humaine.

A cela s'ajoute l'impossibilité structurelle pour l'IA de traiter des conflits de normes qui l'oblige à chercher des corrélations "absurdes" car ajustées de manière forcée (*overfitting*, en anglais) pour expliquer des décisions apparemment conflictuelles, et éviter ainsi la mise en conflit possible de normes entre elles. Tout cela contribue à éloigner d'autant plus la "représentation" du monde et du droit entre l'IA et l'humain.

c. Absence de motivation

La motivation traditionnellement définie est "l'ensemble des motifs d'un jugement ou d'une décision"[18] — le motif étant la raison de fait ou de droit qui commande la décision et que le jugement

[17] La "fiction" est utilisée ici peut être de manière abusive : ce n'est qu'une mise en cohérence, mais il n'y a pas de sens donné au différents éléments.
[18] G. Cornu, *vocabulaire juridique,* 12ème mise à jour, 2018, PUF, p. 671

doit exposer avant le dispositif (CPC, article 455) : il s'agit de "raisons (nécessaires ou surabondantes, exactes ou erronées, suffisantes ou non) que le juge indique comme l'ayant déterminé à se prononcer comme il l'a fait." [19]

La motivation implique donc une intentionnalité (appartenant au registre du "devoir être") qui est étrangère au fonctionnement du système ML quand il juge. En effet, la motivation d'une décision juridique doit être justifiée par (a) l'appréciation des faits sous la forme de causalité, et (b) l'application des normes générales par rapport à ces faits. Le ML ne peut réaliser ni l'une ni l'autre de ces opérations, puisqu'il ne raisonne que par corrélations. S'il est toutefois envisageable de parvenir à identifier le système de raisonnement (explication) ayant abouti à la proposition finale du ML , ce processus explicatif ne peut être qualifié pour autant de motivation au sens juridique.

Une définition plus large de motif—adaptée à l'émergence du *ius artificiale*—pourrait être l'*ensemble des éléments de fait et de droit que le juge indique comme l'ayant déterminé à prononcer la décision comme il l'a fait*. Ainsi, dans la mesure où l'humain pourrait accéder au parcours de fonctionnement de l'IA ayant abouti à sa proposition finale, et dans la mesure où ce parcours s'appuie sur un ensemble d'éléments de faits et de droit (bien qu'ils relèvent de la conceptualisation du monde et du droit qui est propre à l'IA), il serait possible de considérer ici que l'IA mobilise un nouveau type de "motifs" fondé spécifiquement sur des corrélations. Le *ius artificiale* pourrait alors produire des décisions "motivées" par des corrélations.

Pourtant, même dans ce cas, l'intégration du *ius artificiale* au sein du droit positif resterait problématique. Tout d'abord, le procédé de corrélation suppose une mise en cohérence artificielle (potentiellement décorrélée de la raison humaine) des faits et des normes de droit mobilisés par l'IA pour motiver sa décision. Or, cette absence de rationalité (due à l'ajustement exagérément forcé des corrélations) ne relève pas des critères admissibles de l'Etat de droit[20]. De plus, dans la mesure où les éléments de fait et de droit relèvent d'une conceptualisation aliène, étrangère et inaccessible à l'humain, cette motivation —aussi légitime qu'elle puisse être— ne pourra pas être intégrée au sein du droit positif humain.

Par conséquent, il résulte de ces points que la confusion entre l'*être* et le *devoir être*, accompagnée de l'incapacité à identifier des liens de causalité entre les faits à appliquer et les normes juridiques telles qu'elles apparaissent dans le droit positif rend impossible la création d'une décision motivée de la part du ML. Cela implique (1) que le juge-humain ne pourra pas évaluer la compatibilité ou l'incompatibilité des décisions de l'IA avec le système de droit positif—il n'y a donc pas de

[19] Ibid, pp. 670 et 671

[20] Voir O. Pfersmann, « Prolégomènes pour une théorie normativiste de l''Etat de droit ''in : Olivier Jouanjan (dir.), Figures de l'Etat de droit. Le Rechtsstaat dans l'histoire intellectuelle et constitutionnelle de l'Allemagne, Presses Universitaires de Strasbourg 2001, pp. 53-78.

supervision humaine possible sur le ML ; (2) que le juge-humain ne pourra pas se servir de cette décision comme une jurisprudence, c'est-à-dire comme solution éclairante pour les cas à venir (ni sur les faits, ni sur les normes générées).

Ainsi, tant du point de vue de la forme (mode de production du droit) que du fond (contenu du droit produit), le *ius artificiale* est essentiellement un droit a-humain : il est à la fois étranger au système juridique traditionnel et inaccessible à l'humain, et ne peut donc pas s'intégrer dans le droit positif humain.

Conclusion générale :

On ne s'est pas interrogées dans cet article sur les conditions ontologiques nécessaires à la fonction même de juger[21] mais sur les conditions nécessaires à la compatibilité entre une décision humaine et une décision effectuée par des systèmes d'intelligence artificielle. Nous postulons que ces conditions sont liées à l'intériorité du "juge" —que ce soit celle du juge humain qui dispose d'une vie psychique, ou celle de la machine à travers sa boîte noire et son mode de traitement spécifique des informations. Tous deux appréhendent les éléments nécessaires à l'élaboration d'une décision de manière fondamentalement opaque: d'un côté, le "for intérieur" ou "l'intime conviction" du juge humain, qui renvoie à certains aspects de son "intériorité psychique", reste obscure et inaccessible à l'observateur extérieur ; de l'autre côté, la boîte noire de l'IA, dont la fiction est censée simuler celle d'un "juge-type" pour le SE, et celle d'un "juge-moyen" pour le ML. Ces deux typologies d'intériorité (humaine et artificielle) ont des caractéristiques communes lorsqu'elles se manifestent dans le droit : l'inaccessibilité relative (ou la non transparence) du mode de traitement qui se manifeste dans l'espace d'un pouvoir discrétionnaire attribué au juge, et la construction d'une fiction normative (même si, dans le cas de l'IA, cette fiction est assez peu fidèle ou totalement étrangère à la réalité naturelle partagée par les humains). Par ailleurs, ces deux intériorités—humaine et artificielle—ne sont pas du même registre et relèvent de deux modes d'existence différents.[22] La thèse que nous défendons ici est qu'il existe une impossibilité structurelle, c'est-à-dire ontologique, de compatibilité entre la norme juridique issue d'un procédé humain, et la norme juridique issue de l'IA. Cette impossibilité est relative aux différents modes d'existence des intériorités spécifiques et fondamentalement distinctes de l'humain et de la machine. Nous affirmons que, l'une des conditions fondamentales pour que les normes juridiques (humaines et artificielles) soient compatibles les unes avec les autres est l'adhésion des auteurs et destinataires des normes juridiques à une fiction commune du droit et du réel.[23] Cela comprend l'usage de concepts communs renvoyant à des phénomènes

[21] Ainsi il n'a pas été question ici de s'interroger sur les conditions ontologiques (notamment cognitives) pour qu'une entité puisse être auteur d'une norme

[22] Au sens phénoménologique. Voir M. Benasayag, *La singularité du vivant.* Le pommier, 2018

connus et identifiables[24] par le juge qui élabore cette décision, ainsi que la possibilité d'une interprétation du droit —c'est-à-dire la détermination de la signification d'énoncés normatifs—et de l'appréciation des faits du cas d'espèce. Or, les concepts mobilisés par l'IA, notamment dans le cas des systèmes de ML, appartiennent à un monde qu'un humain ne peut pas vraiment se représenter. La mise en fiction du monde par l'IA est soit limitée (pour le SE) soit totalement étrangère à celle du juge humain (pour le ML) de telle sorte que le *ius artificiale* apparaît comme structurellement incompatible avec le droit produit par l'humain.

Par ailleurs, l'usage de l'IA et spécifiquement du ML dans un système juridique soulève des questionnements quant aux conditions de maintien de "l'Etat de droit".[25] En effet le ML produit des solutions jurisprudentielles sans appliquer pour autant une norme déterminée d'un système juridique spécifique. La "décision" rendue par l'IA se fonde sur ce que nous avons appelé une loi "sociologique" des éléments de jurisprudence (comprenant des normes et des faits) fournis à l'IA. Cette "loi" non juridique, élaborée à partir de statistiques et de corrélations effectuées pour l'occasion entre ces éléments de jurisprudence, est censée faire émerger les logiques sous-jacentes du droit jurisprudentiel fourni au système. Or, cette loi ponctuellement élaborée ne vaut que pour le cas d'espèce traité. Elle est en dehors du droit humain et du mode de pensée humaine et ne peut comporter alors aucun élément de motivation. Dans ces conditions, comment envisager encore "d'être le destinataire de normes clairement déterminées"[26] ? L'impossibilité structurelle d'identifier la norme juridique applicable—et a fortiori les conditions de son application—ainsi que l'absence de motivation de la décision issue de certains systèmes d'IA interrogent les conditions de préservation de l'Etat de droit.

Le paradoxe est ici posé : à l'illusion d'un juge agissant en tant que " bouche de la loi", et appliquant mécaniquement—c'est-à-dire sans biais, sans appréciation ni interprétation—une norme juridique, se substituerait aujourd'hui celle d'un juge simple "bouche de l'IA", appliquant aveuglément une norme établie par l'IA, sans possibilité ni de la vérifier, ni de l'interpréter, ni de la motiver.

---

[23] Pour une approche concernant faisant de la culture et du récit la spécificité humaine par rapport à l'IA voir A. Grinbaum, "Placer le numérique au sein de notre histoire", avis du CNPEN, " L'éthique des chatbots", à paraître.

[24] En ce sens on pourrait parler de la nécessité de partager un langage commun, cf J. Bouveresse, *Le mythe de l'intériorité*, Edition de minuit, 1976

[25] entendu ici non comme un "Etat juste" mais comme un Etat qui " soumet le fait au droit" en réduisant la violence au moyen de l'institution de procédures, ce qui implique " un droit fondamental qui est celui d'être le destinataire de normes clairement déterminées (dans la mesure du possible) ».Voir O.Pfersmann, « Prolégomènes pour une théorie normativiste de l'''État de droit'' », op. Cit., p. 77.

[26] Ibid., p. 77. Un autre droit fondamental doit être respecté: « celui de pouvoir fût- ce à titre subsidiaire, demander le contrôle de la conformité des normes par rapport à celles de rang supérieur »,