06 _ 2021

RAPPORT

Mécanisme d'une justice algorithmisée

_Adrien Basdevant
_Aurélie Jean
_Victor Storchan

Coup Data
coupdata.fr

Fondation Jean Jaurès
ÉDITIONS

Adrien Basdevant est avocat au Barreau de Paris. Fondateur d'un cabinet dédié à l'innovation et aux nouvelles technologies, il est aussi le créateur du média en ligne Coup Data, qui prolonge ses travaux autour de la défense des libertés à l'ère du numérique. Membre du Conseil national du numérique (CNNum), Adrien Basdevant est membre du Comité scientifique du département Humanisme numérique du Collège des Bernardins. Il enseigne la protection des données personnelles, l'intelligence artificielle et la cybercriminalité au sein du master ESSEC-Centrale Supélec. Il est co-auteur, avec Jean-Pierre Mignard, de *L'Empire des données. Essai sur la société, les algorithmes et la loi* (Don Quichotte, 2018).

Aurélie Jean, docteure en sciences et entrepreneuse, partage son temps entre le conseil, la recherche et l'enseignement supérieur, principalement à la Sloan du Massachussets Institute of Technology. Elle est présidente et fondatrice de l'agence In Silico Veritas, spécialisée dans le développement algorithmique, et *Chief AI Officer* et cofondatrice de DpeeX, une startup *deeptech* dans le domaine de la médecine de précision pour le cancer du sein, utilisant de l'intelligence artificielle. Elle est l'auteure, entre autres, de deux ouvrages parus aux éditions de l'Observatoire : *De l'autre côté de la Machine. Voyage d'une scientifique au pays des algorithmes* (2019) et *L'Apprentissage fait la force* (2020).

Victor Storchan est ingénieur dans l'industrie financière et contributeur régulier à Phébé, rubrique de veille d'articles de recherche du magazine *Le Point*. Il est ancien élève de l'université de Stanford et de l'ENS Lyon.

Introduction

Ce rapport a pour objectif de décrire, sous un angle scientifique et juridique, les grands principes de la justice algorithmisée, ainsi que ses mécanismes sous-jacents. Il convient de préciser que l'adjectif « algorithmisée » est ici préféré à celui de « prédictive », car il exprime plus justement le fonctionnement algorithmique de cette justice loin de toute prédiction *stricto sensu*. Ce rapport se concentre sur les applications d'une approche algorithmisée de la justice en fonction du domaine et des acteurs (avocat, juge, justiciable, etc.) considérés. Il se veut également interdisciplinaire et pédagogique, afin d'être lu et compris par tous, notamment par les citoyens, premiers concernés par son déploiement.

Ce rapport est structuré en six parties. La première et la deuxième partie visent à présenter l'histoire et la relative pertinence d'une justice algorithmisée, ainsi que les définitions des notions techniques utilisées, telles que les algorithmes, la *data* ou encore l'intelligence artificielle (IA). La troisième partie se concentre sur les différents procédés de justice algorithmisée en fonction du type de système judiciaire. La quatrième partie est dédiée à l'étude de cas concrets, mis en place en Europe et aux États-Unis. L'étude du fonctionnement algorithmique sous-jacent à chacun de ces cas permet de révéler le potentiel et les limites de telles solutions. La cinquième partie cherche à définir les précautions et les écueils inhérents au développement algorithmique en matière judiciaire. Enfin, la sixième et dernière partie porte sur quelques recommandations.

Ce travail de définition des termes et des concepts est un préalable nécessaire au développement constructif et responsable des technologies pour les domaines de la justice. Ce rapport cherche plus particulièrement à identifier les solutions algorithmiques applicables en matière de justice pour des raisons juridiques et/ou techniques. Un état de l'art en sciences algorithmiques sera également dressé afin de lister les avancées qui impacteront le déploiement de solutions actuellement envisagées dans la justice.

Ce rapport ne présente pas une liste exhaustive de tous les cas pratiques et de toutes les recherches réalisées dans le domaine de la justice algorithmisée, mais s'intéresse aux actes de juger et de décider concernant un prévenu ou un individu. Elle est un point d'ancrage d'une réflexion sur les tenants et les aboutissants d'une telle mécanique judiciaire ; et ce pour tous les corps de métier en lien avec la justice, mais aussi les sciences. Ce rapport constitue également une sorte de grille de lecture pour le citoyen, afin qu'il capture les mécanismes de ce nouveau tissu judiciaire dont il est directement le sujet.

Histoire et pertinence d'une justice algorithmisée

Les origines de la justice prédictive

La tentation de vouloir mettre en équation l'ensemble des interactions sociales pour prédire leurs survenances et gouverner leurs effets n'est pas nouvelle. En 1706, Leibniz, que l'on présente comme un des pères du calcul infinitésimal (une des briques mathématiques essentielles pour l'IA) écrit dans *Opinion sur les principes de Pufendorf*[1] :

> Ni la norme de bonne conduite, ni l'essence du juste ne dépendent de la libre décision [de Dieu], mais plutôt des vérités éternelles, objets de l'intellect divin, qui constituent, pour ainsi dire, l'essence de la divinité elle-même […]. Et, en effet, la justice suit certaines règles d'égalité et de proportion [qui ne sont] pas moins fondées dans la nature immuable des choses et dans les idées divines que ne le sont les principes de l'arithmétique et de la géométrie.

Leibniz illustre ici combien il est tentant de vouloir utiliser la logique algorithmique dans le but de parfaire l'exercice de la justice. Trois ans plus tard, Nicolas Bernoulli applique les découvertes de son oncle, le probabiliste Jacques Bernoulli, au domaine du droit et s'emploie à défendre l'usage des algorithmes dans le champ juridique. Ainsi, dans sa thèse *De Usu Artis Conjectandi in Jure* obtenue à l'université de Bâle, il se pose la question suivante : quelle est la pertinence de l'utilisation des probabilités pour résoudre les questions que soulève le droit ? Selon l'analyse de l'historien du droit Robert Carvais[2], Nicolas Bernoulli s'appuie sur le droit romain pour proposer des applications juridiques de la notion de probable, notamment pour les questions d'attribution de pensions alimentaires ou l'achat des rentes viagères. Cette thèse inspira de nombreux scientifiques, tels que Condorcet (*Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, 1785), Laplace (*Essai philosophique sur les probabilités*, 1795) ou encore Poisson (*Recherches sur la probabilité des jugements en matière criminelle et en matière civile*, 1837), qui poursuivirent l'exploration de l'application d'un cadre mathématique au domaine juridique.

La première polémique publique sur l'utilisation des statistiques apparaît lors du débat sur l'inoculation de la variole. La controverse s'ouvrit lorsqu'en 1774 la variole emporta le roi, conduisant son successeur, Louis XVI, à inoculer le vaccin à l'ensemble de la famille royale. La question se posait alors de savoir si l'hygiène publique serait davantage assurée en rendant obligatoire ce vaccin.

Pour le juriste Daniel Bernoulli, les « chances de gain » militent en faveur d'une campagne de vaccination. Il propose d'appliquer une formule similaire à celle utilisée dans les jeux de hasard pour résoudre la question politique de la vaccination. Il calcula ainsi que les chances de gain, en l'occurrence une espérance de vie plus grande de trois ans pour les individus inoculés, permettaient de conclure à l'utilité de la vaccination.

1. Gottfried Wilhelm Leibniz, *Le Droit de la raison*, voir « Monita quaedam ad Samuelis Puffendorfii Principia », 1706.
2. Robert Carvais, « Anticipation et réception d'une thèse de droit "*De Usu Artis Conjectandi in jure*" de Nicolas Bernoulli », *Journal électronique d'Histoire des probabilités et de la statistique*, vol. 2, n°1, juin 2006.

À l'inverse, le philosophe et mathématicien français Jean Le Rond d'Alembert et le médecin français Claude Bernard s'opposaient à cette approche, qui entraînait à leurs yeux une confusion entre la moyenne et la norme. Selon eux, il ne devrait pas exister d'équivalence entre le fait observé et le droit qui en découle. Cette opposition donne finalement raison aux tenants de la vaccination. Elle constitue une des premières occurrences de la victoire du calcul statistique sur la loi.

Ainsi, l'idée de s'appuyer sur la machine pour rendre justice précède largement les débuts de l'IA. Près de deux siècles plus tard, en 1949, Lee Loevinger[3], juriste américain, introduit le terme de « *jurimetrics* » pour désigner l'utilisation des corrélations statistiques et des modèles probabilistes dans la justice, en proposant une liste d'applications pour répondre à des questions légales concernant les juges, les témoins ou le législateur.

Naissance (maladroite) de la formule « justice prédictive »

La « justice prédictive » nourrit le fantasme d'une justice automatique, fluide, efficace, débarrassée des inconvénients humains et qui pourrait à terme renforcer la confiance des justiciables dans la justice.

Pour autant, la formule « justice prédictive » est imprécise. Elle l'est dans la mesure où elle repose sur la croyance que l'utilisation d'outils algorithmiques analysant la jurisprudence existante permettrait de prédire ce que sera la jurisprudence future. Cette croyance est erronée. D'une part, parce que le passé ne permet pas nécessairement de prédire l'avenir. D'autre part, parce que recourir à l'apprentissage automatique ne permet pas pour autant de connaître l'argumentation juridique fondant une décision. On comprend donc que la « justice prédictive » aide à quantifier l'aléa plutôt qu'à prédire le contenu d'une décision.

La formule « justice prédictive » contribue également à tromper notre perception du rôle de la justice. L'objectif poursuivi n'est pas tant de rationaliser la décision judiciaire, pour la rendre plus prédictible, et donc plus sûre. Autrement dit, le but n'est pas d'augmenter la sécurité juridique en faisant disparaître l'aléa, car le propre de la justice n'est pas l'uniformisation, mais, au contraire, l'ajustement à chaque cas, considéré comme singulier. La réponse ne peut et ne doit pas être homogène.

Affirmer que la justice peut être prédite ou prédictive revient à confondre le fait (ce qui est) et le droit (ce qui doit être). Le droit n'a pas pour objet de saisir directement le monde des faits. La différence entre le fait et le droit, c'est précisément celle entre l'être et le devoir être. Lorsque les juristes étudient un dossier, ils ont pour habitude de qualifier les faits d'une affaire, afin de leur appliquer les règles de droit correspondantes. Ils opèrent ainsi un lien entre les faits observés et les normes existantes. En voulant mesurer le réel sans l'interpréter, toute qualification est contournée et donc toute interprétation juridique est mise de côté. Au droit, le *Big Data* substitue les faits. S'il est confondu avec la réalité, autrement dénommée vérité, le droit est privé de sa mission, qui est la recherche de compromis entre des intérêts contradictoires d'une société.

À la crise de représentativité contemporaine, à laquelle sont déjà confrontés les États-nations, s'ajoute donc la crise de représentativité du réel par les statistiques. C'est bien l'un des problèmes majeurs auquel nous sommes aujourd'hui confrontés : quelle représentation se fait-on du monde et de nous-mêmes ? Sommes-nous uniquement réduits à notre *data* et appréhendés par une approche quantitative ?

Les algorithmes pourraient conduire à une « factualisation du droit », selon la formule d'Hervé Croze[4]. Cela reviendrait à tout mettre sur le même plan factuel, et donc rendre comme *data* analysable tout type d'information lié à l'affaire : contexte du dossier, jurisprudences passées applicables, précédents jugements rendus par les juges de l'affaire, etc. Cette

3. Lee Loevinger, « Jurimetrics.The Next Step Forward », *Minnesota Law Review,* 1949.
4. Hervé Croze, « Justice prédictive : la factualisation du droit », *La semaine juridique, édition générale*, LexisNexis, janvier 2017.

mise en données du réel permettrait de modéliser le droit, alors même qu'il existe une hiérarchie entre les différents textes de droit (loi, règlement, décret) et que certains aspects ne sont pas nécessairement comparables (jurisprudences passées et position des juges saisis de cette affaire).

Algorithmiquement (informatiquement), ces critères seraient pondérés en amont, mais comment ? Le risque principal serait d'aboutir à une confusion entre le fait et le droit en faisant des caractéristiques déterminantes identifiées par l'algorithme dans la jurisprudence le fondement de la motivation d'un juge.

Or, il n'y a pas équivalence entre les deux. Le fait d'une affaire doit être qualifié et interprété en droit. Le droit est ainsi une opération de qualification juridique des faits. Les faits sont établis par les parties qui argumentent ensuite en droit, mais c'est finalement le juge qui applique le droit. À cet égard, il est intéressant de souligner, comme le rappelle la professeure de droit Cécile Bourreau-Dubois, qu'en matière d'indemnisation, le droit français interdit au juge de fonder la motivation d'une décision exclusivement sur un barème[5]. Ce raisonnement pourrait s'appliquer de la même façon aux algorithmes de « justice prédictive ». En effet, ils ne font que proposer la solution juridique la plus probable au regard de circonstances déterminantes préalablement identifiées dans la jurisprudence, ce qui s'apparente, *in fine*, à une forme de barème[6]. Ainsi, tout juge devrait pouvoir comprendre comment un outil algorithmique d'aide à la décision interprète les faits, sinon il risquerait de se trouver lié par des recommandations qu'il ne saurait pas motiver en fait et en droit.

La question de l'algorithme et la justice n'est pas tant celle de la substitution du juge par la machine, mais davantage celle de l'impact de nouveaux outils d'aide à la décision sur la faculté de juger, de rendre le droit et de prendre des décisions judiciaires. C'est pourquoi nous parlerons dans ce rapport de « justice algorithmisée » davantage que de « justice prédictive ».

Pertinence d'une justice en partie algorithmisée

Le fantasme de remplacer le juge par la machine repose sur l'idée que toute interaction sociale peut être mise en équation et que le raisonnement judiciaire pourrait être automatisé ou mécanisé. Il convient pourtant de revenir aux différentes fonctions du travail judiciaire. Il s'agit non seulement de rendre une décision de justice, mais également de considérer toute la portée symbolique qui y est associée ; le fait de se rendre dans une enceinte judiciaire, d'entendre contradictoirement les arguments des parties prenantes, par exemple. Pour Antoine Garapon et Jean Lassègue, il y a une théâtralité consubstantielle à l'acte judiciaire[7].

En effet, le processus judiciaire implique traditionnellement un tiers impartial (un juge ou des juges), qui symbolise les rapports humains. En organisant un procès (civil, pénal, administratif), chaque partie a un rôle. Cette symbolique, notamment matérialisée par le port de la robe, l'entrée dans un tribunal, le décor, l'ordre de passage, organise et met à distance chacun. Une question s'impose alors : comment sera recréée cette symbolique dans un environnement dématérialisé ? Dans ces circonstances, comment seront pris en compte les émotions, les sensations, les intuitions et les raisonnements de chacun ?

Par ailleurs, certains principes fondamentaux de la procédure requièrent de garantir l'accès à un juge impartial, le droit au procès équitable, l'égalité des armes, l'assistance par un avocat, la présomption d'innocence, ou encore le principe du contradictoire. Existera-t-il, par manque de recul ou de connaissances, un risque de partialité ou de dépendance des juges s'appuyant sur un outil d'aide à la décision algorithmique ? Existera-t-il un risque de rupture d'égalité des armes entre les parties équipées d'outils algorithmiques et celles qui n'en disposent pas ? Les

5. Cass. 1re Civ., 23 octobre 2013, n°12-25.301 dans lequel la Cour de cassation a censuré un arrêt d'appel qui s'était fondé sur une table de référence, annexée à une circulaire, pour déterminer le montant des contributions dues par un père au titre de l'entretien et de l'éducation de son enfant.
6. Cécile Bourreau-Dubois, « La barémisation de la justice : une approche par l'analyse économique du droit », Rapport final, Mission droit et justice, février 2019.
7. Antoine Garapon et Jean Lassègue, *Justice digitale*, Paris, Presses universitaires de France, 2018.

décisions de justice seront-elles suffisamment « motivées » si elles sont rendues par des suggestions algorithmiques dont les logiques de fonctionnement nous échappent ?

Enfin, le jugement ne constitue pas la reproduction probabiliste de cas passés. Il existe, par exemple, des situations de revirement jurisprudentiel, à savoir d'évolution des solutions retenues dans des configurations données. Ainsi, la modélisation, le calcul, la prédiction – aussi sophistiqués puissent-ils être – de cas passés ne peuvent pour autant pas abolir la possibilité d'introduire du nouveau et de l'aléa. Cela reviendrait à s'enfermer dans des raisonnements passés et affirmer une forme de fatalité, niant tout contradictoire, toute présomption d'irréductible singularité des cas jugés et, de fait, l'obligation de rendre une décision individualisée[8].

L'utilisation du *Big Data* ne fera pas disparaître l'incertitude et ne pourra supprimer l'aléa judiciaire, celui-ci n'étant pas uniquement la résultante de biais humains. Telle affirmation permet de s'interroger sur la fonction de l'aléa judiciaire. En effet, même dans l'hypothèse où la technologie serait en mesure de supprimer tout aléa judiciaire, cela serait-il vraiment souhaitable ? La suppression d'un tel aléa ne signerait-elle pas la mort du droit ? Rappelons que le droit est une matière vivante dont le propre est de pouvoir se réinventer pour s'adapter aux évolutions de la société. Cette faculté implique nécessairement l'existence d'un aléa incompressible.

8. Adrien Basdevant et Jean-Pierre Mignard, *L'Empire des données. Essai sur la société, les algorithmes et la loi*, Paris, Don Quichotte, 2018.

Algorithmes, *data* et intelligence artificielle

Introduction à l'intelligence artificielle

Même si le terme semble paradoxal (en quoi l'intelligence peut être artificielle ?), il est aujourd'hui largement utilisé pour exprimer une discipline qui remonte à l'époque des premiers ordinateurs, vers la moitié du XX^e^ siècle[9]. Plus qu'une discipline, l'intelligence artificielle est aussi un outil : un moyen pour résoudre un problème, répondre à une question ou analyser et comprendre des mécanismes.

L'intelligence artificielle regroupe ainsi l'ensemble des méthodes permettant de simuler un phénomène ou un scénario, qu'il soit physique, chimique, médical, sociologique, démographique ou encore juridique. Un modèle est traduit algorithmiquement pour reproduire numériquement, par une simulation sur ordinateur, le phénomène ou le scénario en question.

L'intelligence artificielle regroupe deux types de méthodes algorithmiques : d'une part, les méthodes algorithmiques dites explicites, dans lesquelles l'ensemble de la logique est défini au sein de l'algorithme explicitement par les humains ; d'autre part, les méthodes implicites dans lesquelles cette même logique est décrite implicitement par apprentissage[10] – on parle également de *Machine Learning*. Dans ces deux approches, les données vont permettre d'alimenter l'algorithme, soit par calibration (algorithme explicite), soit par apprentissage (algorithme implicite), afin que ce dernier exprime de manière réaliste et juste le phénomène à simuler.

Comprendre la forme de cette donnée (ou *data*) permet de comprendre l'impact sur l'algorithme ainsi que les simulations et les résultats qui en résultent. De plus, cela permet d'anticiper les futures évolutions et leurs influences possibles sur la prochaine génération de technologies appliquées à la justice algorithmisée.

Data structurée *versus* non structurée

Les *data* correspondent à l'ensemble des informations qui décrivent une situation, un phénomène, ou encore un état. Ces *data* sont, par exemple, des images, des PDF, des contenus audio, des vidéos, des nombres, des graphes, ou encore un mail ou un texte. Parmi les *data*, on distingue les *data* structurées des *data* non structurées.

Les *data* dites non structurées sont les *data* brutes qui ne sont pas mathématisées ou formatées pour pouvoir décrire quantitativement la situation, le phénomène ou l'état en question. Ces *data* sont, par exemple, un document PDF, une photo, ou encore un mail. *A contrario*, les *data* structurées sont formatées, voire mathématisées, pour traduire quantitativement leurs contenus. Une photo peut être décrite mathématiquement comme une matrice de nombres décrivant la couleur de chaque pixel, par exemple par un vecteur de dimension 3 (RGB) encodant les couleurs primaires, ou comme un spectre de fréquences.

9. Marvin Minsky, « Steps Toward Artificial Intelligence », *Proceeding of the IRE*, vol. 49, n°1, janvier 1961.
10. Aurélie Jean, « Une brève introduction à l'intelligence artificielle », *Médecine & Sciences*, vol. 36, n°11, novembre 2020, pp.1059-1067.

Ces deux types de données peuvent être labellisés : cela signifie que pour une tâche donnée (classification, régression, etc.), on associe le résultat de la tâche (catégorie, score, etc.) à chaque donnée, afin que le modèle puisse apprendre les corrélations entre données d'entrée et résultats. Cette étape d'annotation est généralement chronophage, car le volume de données à labelliser est important. Selon la tâche à effectuer, cette étape nécessite le concours d'experts (dans le cas de préparation de données de segmentation d'images de tumeur, par exemple), ou peut être complètement externalisée, grâce à des outils collaboratifs qui mobilisent des non-experts chargés d'annoter les données suivant les instructions précises données par les ingénieurs.

Ainsi, pour trier les dizaines de millions d'images du premier jeu de données massif utilisé pour la classification d'images appelé ImageNet, la chercheuse Li Fei-Fei[11] a fait appel à plus de 6 000 personnes provenant de plus de 150 pays[12]. De même, dans le cadre d'un programme de l'université de Berkeley, des doctorants en droit ont été récemment mobilisés pour annoter 10 000 pages de contrats juridiques après avoir reçu 100 heures de formation pour un coût total de 2 millions de dollars[13].

De ce fait, cette étape est un frein majeur à l'adoption de l'IA dans la plupart des industries ou domaines d'application, dont la justice (voir la partie « Enjeux technologiques et scientifiques dans la justice algorithmisée »). La recherche s'emploie à inventer des moyens, comme les techniques de *self supervised learning*[14], pour l'automatiser. Enfin, dans le cas des données non structurées, il est souvent nécessaire en amont de calculer des *features* (ou ensembles de critères) sur lesquels le modèle apprend à effectuer la tâche. Il est, par exemple, possible d'extraire d'un mail un ensemble de *metadata* : la date d'envoi, le nom de l'expéditeur et celui du destinataire, ou encore la tonalité du texte. Autrement dit, une *metadata* est une donnée servant à en décrire une autre.

Aujourd'hui, les *data* structurées sont stockées au sein de bases de données traditionnelles, tandis que les *data* non structurées le sont au sein de lacs de données (*data lake*).

Stocker les *data* non structurées évite de réfléchir à la description quantitative de ces *data* et permet donc d'élargir leur utilisation. À l'inverse, les *data* structurées sont exploitables directement par un algorithme. La dématérialisation de l'information (numérisation de documents papiers, par exemple), l'émergence de capteurs d'information et de l'IoT (*Internet of Things*), ou encore l'augmentation des puissances de calculs ont permis des avancées dans le traitement algorithmique associé, et donc l'élargissement des problèmes à résoudre.

Algorithmes explicites *versus* implicites

Un algorithme est constitué d'un ensemble d'opérations exécutées selon une certaine logique et une certaine hiérarchie. Même si la science algorithmique existe depuis plus de deux mille ans, on fait référence aujourd'hui à l'algorithme numérique qui est destiné à être implémenté dans un code de calcul pour tourner sur un ordinateur. Alors que les algorithmes traditionnels étaient exécutés à la main, les algorithmes numériques sont opérés par des programmes.

Les opérations, ainsi que la logique associée de ces algorithmes numériques, sont explicitement ou implicitement définies et réalisées. Comme introduit précédemment, on parle aussi d'algorithmes explicites et d'algorithmes implicites.

Les algorithmes explicites, comme leur nom l'indique, sont décrits explicitement par les concepteurs (scientifiques ou ingénieurs) à travers l'écriture de conditions, d'hypothèses, voire d'équations mathématiques.

11. J. Deng, W. Dong, R. Socher, Li-Jia Li, Kai Li, Li Fei-Fei, « Imagenet: A large-scale hierarchical image database », *IEEE conference on computer vision and pattern recognition*, 2009, p. 248-55.
12. À écouter sur Listen Note, Li Fei-Fei, épisode 44, 8 juillet 2020.
13. Dan Hendrycks, Collin Burns, Anya Chen et Spencer Ball, « CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review », arXiv preprint, 2021.
14. Voir Kyle Wiggers, « Yann LeCun and Yoshua Bengio : Self-supervised learning is the key to human-level intelligence », *Venture Beat,* 2 mai 2020.

Les *data* sont alors utilisées pour calibrer l'algorithme, pour permettre d'identifier les constantes de possibles équations traduisant le phénomène ou le scénario à simuler. Ces *data* servent également de données d'entrées sur lesquelles l'algorithme est exécuté.

À titre d'exemple, dans les travaux en justice pénale de Sharad Goel et ses collaborateurs[15], un algorithme basé sur des règles explicites utilise des critères juridiques pertinents et co-sélectionnés par les juges eux-mêmes. Cet algorithme assiste les juges dans leur prise de décision quant au placement en détention provisoire ou à la remise en liberté, précédant le jugement. En pratique, l'algorithme, qui s'appuie sur l'expertise et l'expérience des juristes, estime les risques de défaut de comparution des accusés en fonction de plusieurs critères : le type de charges retenues contre le prévenu, le type de stratégie de défense adoptée par l'avocat, ou encore son historique de défauts de comparution.

Les algorithmes implicites sont, quant à eux, décrits implicitement lors de la phase dite d'entraînement sur des *data* dites d'apprentissage. L'algorithme se construit sur la résolution d'un problème d'optimisation constituant l'apprentissage, à partir de données d'entrée, représentant l'ensemble des scénarios et des situations possibles, et des résultats supposés.

Dans les mêmes travaux cités dans le paragraphe précédent, Sharad Goel et son équipe entraînent des modèles statistiques implicites[16] sur 165 000 cas de comparutions passées devant une juridiction pénale des États-Unis. Ces résultats sont ensuite analysés selon une approche contrefactuelle, afin de mettre en évidence des relations implicites de causalité. Plus précisément, l'utilisation de techniques d'inférence causale, comme le *propensity score*, permet de reproduire fidèlement le processus des essais rendus aléatoires (approche de référence pour évaluer les effets des traitements ou des interventions sur des résultats).

Ces modèles implicites sont plus efficaces que les modèles basés sur des critères judiciaires explicitement construits en collaboration avec les juges. Cela pourrait provenir des variabilités entre les décisions des juges sur la libération conditionnelle : la faible corrélation entre le risque de fuite et la décision de libération conditionnelle est en partie attribuable à des différences notables dans les taux de mise en liberté parmi les juges, certains libérant plus de 90 % des accusés et d'autres n'en libérant que 50 %.

Sans entrer dans les détails des différents types d'apprentissage, on différencie tout de même l'apprentissage supervisé de l'apprentissage non supervisé[17]. Alors que l'apprentissage supervisé, couramment utilisé, est réalisé sur un ensemble de données labellisées, l'apprentissage non supervisé procède sur un ensemble de données non labellisées. Si l'apprentissage non supervisé est un jour généralisé à tous les types de problèmes, cela serait un bond technologique important dans l'accélération des apprentissages algorithmiques, car la phase très chronophage de labellisation des données serait écartée.

Dans ces deux ensembles d'algorithmes, il existe de profondes différences dans la manière de résoudre *stricto sensu* un problème, dans le niveau d'explicabilité et d'interprétabilité algorithmique, ou encore dans la valeur ajoutée du résultat algorithmique quant à la compréhension d'un contexte ou d'une situation. Par sa plus forte abstraction, l'algorithme implicite est capable de résoudre un problème sur lequel on connaît peu de choses ou un problème difficilement mathématisable. Par conséquent, dans de nombreux cas, les algorithmes implicites sont bien plus efficaces que les algorithmes explicites.

Cela étant dit, par son caractère ambigu, l'algorithme implicite est plus difficile à expliquer et à interpréter, contrairement à l'algorithme explicite dont on connaît toutes les étapes de résolution et sa logique sous-jacente. Les algorithmes implicites sur un apprentissage non supervisé sont plus difficilement explicables dans la mesure où les *data* sur lesquelles ils s'entraînent ne sont pas labellisées. Autrement dit, on ne sait

15. Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel et Daniel G. Goldstein, « Simple rules to guide expert classifications », *Journal of the Royal Society, Statistics in Society Series A*, 27 mai 2020.
16. De type régression logistique et forêt aléatoire.
17. Yann Le Cun, *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*, Paris, Odile Jacob, 2019.

pas sur quelle *metadata* ces algorithmes apprennent. Du fait du plus haut niveau d'explicabilité et d'interprétabilité des algorithmes explicites, il est plus fréquent et quasi systématique d'analyser leurs résultats, dans le but de comprendre les mécanismes du phénomène simulé. Néanmoins, certains algorithmes implicites d'apprentissage statistique sont plus explicables, ceux utilisant la régression linéaire, par exemple.

Comme souligné lors de la dernière conférence NeurIPS[18] de 2019, relever les grands défis de nos temps, tel que le changement climatique, impose de considérer comme complémentaires les approches explicite et implicite dans les modèles d'IA actuellement développés. Sans une compréhension fine des mécanismes des phénomènes simulés, il sera difficile de trouver une solution motivée, efficace et durable.

Enjeux technologiques et scientifiques dans la justice algorithmisée

Les développements technologiques en IA reposent sur trois conditions : la mise à disposition de volumes de données, une puissance de calcul suffisante et l'implémentation et l'usage d'algorithmes performants. Ces trois conditions soulèvent des enjeux technologiques et scientifiques fondamentaux pour le déploiement et l'usage d'outils dans la justice algorithmisée.

La collecte de données requiert leur dématérialisation et leur traitement numérique pour en extraire l'information pertinente. En France, plusieurs initiatives ont vu le jour pour centraliser et référencer plusieurs millions de décisions judiciaires, telles que JuriCa (accès toutefois payant pour les éditeurs juridiques) ou Jurinet pour les décisions d'appel, Légifrance (textes législatifs, décisions de juridictions supérieures) ou encore Ariane (jurisprudence administrative). Ces initiatives doivent se généraliser pour permettre une diffusion plus large des données, tout en garantissant un niveau de confidentialité et de confiance indispensable entre les acteurs.

Le rapport Bothorel sur la politique publique de la donnée, remis en décembre 2020 au Premier ministre[19], étudie notamment l'utilisation par la plateforme Infogreffe, qui centralise les données collectées par les greffes des tribunaux de commerce. Le rapport souligne que la mise à disposition de ces données de justice commerciale en accès libre est loin d'être achevée. L'étape d'automatisation de l'anonymisation de ces contentieux juridiques constitue le principal frein à l'accès simple et systématique à ces données. Ces *data* sont pourtant identifiées comme étant un jeu de données « à forte valeur » pour les startups. Combinées aux actions des entreprises en *Legal Tech*, l'identification des meilleures pratiques (rapport d'impacts, accompagnement des producteurs de données, amélioration de l'interopérabilité des données) participe à la démocratisation des pratiques d'*open data*. En effet, cette mise à disposition libre et gratuite des données judiciaires pseudonymisées et conformes à la législation européenne introduit davantage de transparence dans les décisions judiciaires : les procédures d'audit sont facilitées.

Comme l'expose clairement la mission d'étude et de préfiguration sur l'ouverture au public des décisions de justice menée par Loïc Cadiet, les pratiques d'*open data* sont aussi un levier d'accélération dans la recherche académique ainsi que dans la recherche et développement en entreprise. Elles facilitent, en effet, l'accès aux données nécessaires à l'élaboration des modèles d'IA dans le domaine de la justice[20].

Parmi les nombreuses initiatives peuvent être notamment citées :

– en France, sous l'impulsion de la loi de 2016 pour une république numérique[21] et de l'initiative Open

18. Conference on Neural Information Processing Systems.
19. « Remise du rapport sur la politique publique de la donnée, des algorithme et des codes sources », Gouvernement, 23 décembre 2020.
20. Loïc Cadiet, « L'*open data* des décisions de justice », Mission d'étude et de préfiguration sur l'ouverture au public des décisions de justice, novembre 2017.
21. Voir « La loi pour la République numérique », ministère de l'Économie, des finances et de la relance.

Justice[22] utilisant les bases de données de Jurinet et de JuriCa, 1 500 décisions de justice annuelles (soit plus de 8 % de l'ensemble des décisions judiciaires) sont rassemblées et mises à disposition librement et gratuitement ;

– en Californie aux États-Unis, une plateforme d'*open data*[23] rassemble les décisions judiciaires pour assister la recherche de jurisprudence, mais aussi les activités des avocats par branches (corruption, criminalité organisée, droit civil, etc.), dans un but de transparence de la justice et de collaboration de ses acteurs.

Le déploiement de plateforme d'*open data* permet de satisfaire le besoin de données en quantité, mais aussi en diversité, afin d'entraîner et/ou de calibrer les algorithmes.

Un algorithme, qu'il soit explicite ou implicite, sera considéré comme robuste s'il est capable, entre autres, de simuler correctement une réponse dans un contexte différent et avec des données différentes de celles fournies pour le calibrer ou l'entraîner. En revanche, même robuste, l'algorithme ne saurait être considéré comme un outil de prise de décision, au moins pour la matière judiciaire. Pour formuler ses résultats, l'algorithme, implicite ou explicite, se base toujours sur l'existant, et donc le passé. En droit, « l'existant » est constitué par la jurisprudence, c'est-à-dire des décisions similaires au cas considéré.

Ce type de raisonnement par analogie, même efficace en général, a ses limites. L'algorithme initialement décrit ou entraîné peut voir son domaine d'application évoluer, et donc son efficacité simulatoire modifiée. Dans le domaine de la justice, des changements dans la législation ou dans la jurisprudence peuvent diminuer la robustesse de l'algorithme s'il n'a pas été conçu en conséquence. Une adaptabilité et un suivi du comportement des algorithmes sont alors requis.

Dans le cas des algorithmes d'apprentissage supervisé, l'entraînement est réalisé sur des données structurées qui nécessitent une intervention humaine importante pour identifier et labelliser chaque point de données (ou *data* point). Dans le cas d'un apprentissage profond supervisé, il faut un très grand nombre de données. Par conséquent, la durée de traitement de ces données est pénalisante dans la mesure où elle pèse sur le temps de développement et de déploiement d'une technologie, voire tout simplement dans la capacité à la développer.

De nombreux travaux de recherche s'intéressent aux techniques non supervisées. Ces dernières permettraient d'entraîner un algorithme sur des données non labellisées. De nouvelles techniques de transfert de connaissances (*transfer learning*) ou de résolution multitâches (*meta learning*) apparaissent ainsi prometteuses. Dans le premier cas, on transfère la connaissance d'un système, ayant appris sur des données massivement disponibles, à un autre. Ce transfert permet d'initialiser ce second système et de l'affiner à l'aide d'un jeu de données bien moins volumineux. Dans le second cas, on utilise la connaissance de systèmes ayant appris des tâches simples pour effectuer des tâches plus compliquées. L'idée est que la connaissance puisse être généralisable à des tâches inédites. Ainsi, en langage naturel où les données labellisées sont rares, des systèmes peuvent conjointement être entraînés à classer des documents par thèmes avant d'appliquer des algorithmes d'entités nommées (identification de lieux, noms propres, personnes, entreprises, etc.).

Un enjeu critique, prégnant dans le domaine de la justice, concerne l'explicabilité des algorithmes. La justice doit pouvoir rendre compte des raisons et des motivations des décisions judiciaires, et s'assurer de la compréhension et de l'application de la loi par et pour tous. L'usage d'algorithmes dans l'aide à la décision ou dans certaines procédures judiciaires exige de pouvoir expliquer leur fonctionnement et leurs comportements selon les données d'entrée. Les algorithmes explicites sont par définition entièrement explicables dans la mesure où ils sont décrits entièrement explicitement par les concepteurs. Les algorithmes d'apprentissage sont plus difficilement explicables du fait de la description implicite de leur logique.

22. Voir « Open Justice », Entrepreneurs d'intérêt général.
23. Voir « Open data » sur le site du département de la Justice des États-Unis.

Aujourd'hui, on est capable d'expliquer, ou *a minima* d'interpréter, le fonctionnement d'un algorithme par une description précise des données d'entraînement et par une analyse des réponses de l'algorithme. Les enjeux de l'explicabilité des algorithmes ne résident pas tant dans la faculté d'expliquer le fonctionnement de l'algorithme que dans celle de pouvoir expliquer les résultats rendus. Cela implique de pouvoir identifier quels types de circonstances ont été pris en compte par l'algorithme et quelle a été leur pondération. Dans les prochaines générations d'algorithmes d'entraînement non supervisés, un travail supplémentaire devra donc être réalisé afin d'expliquer et d'interpréter ce sur quoi les algorithmes apprennent. En effet, les *data* n'étant pas labellisées, il sera plus difficile de décrire sur quelle quantité analytique l'algorithme est entraîné.

Que ce soit dans la collecte et le traitement des données, ou dans le développement et l'usage d'algorithmes toujours plus performants, des enjeux éthiques, de transparence et de responsabilité s'imposent pour permettre le déploiement des technologies d'IA dans le domaine de la justice. Une coopération de confiance entre les acteurs de la justice et ceux de la technologie, des sciences de la *data* et de l'IA, permettra la co-construction des outils efficaces n'induisant aucun traitement inégal et injuste des individus. Dans cette perspective, l'analyse de la présence possible de biais algorithmiques pouvant mener à la discrimination technologique doit être menée[24]. À ce titre, l'Institut Montaigne propose dans un récent rapport[25], réalisé en collaboration avec des chercheurs et des ingénieurs du domaine de l'IA, quelques solutions techniques et non techniques pour lutter contre ces biais.

24. Florence G'sell (dir.), *Le Big Data et le Droit*, Paris, Dalloz, 2020.
25. « Algorithmes : contrôle des biais S.V.P. », Institut Montaigne, 2020.

La justice algorithmisée selon les systèmes judiciaires

En règle générale, l'IA se base sur les événements passés pour inférer des prédictions futures. Pour cela, elle identifie, dans un large volume de données historiques, les corrélations pertinentes. Dans le cas de la justice algorithmisée, ces données sont en majeure partie textuelles et représentent des cas de litiges. En matière judiciaire, les outils algorithmiques ont pour objectif d'aider le juge dans sa prise de décision. La stratégie pour y parvenir peut s'avérer très différente d'un système juridique à un autre, et donc d'un pays dans lequel sont déployés ces outils à l'autre.

On distingue principalement deux systèmes : la *common law* et le droit civil. Tous deux ont des implications directes sur la manière dont on peut envisager d'entraîner les modèles d'IA pour la justice algorithmisée.

La justice dans les pays du *common law*

Le système juridique de *common law* s'est développé dans les pays anglo-saxons et les colonies anglaises, il est notamment appliqué aujourd'hui aux États-Unis, au Royaume-Uni, en Australie, au Canada ou encore en Inde. Il s'agit d'un régime juridique non écrit, qui recourt très peu aux normes textuellement consacrées et se fonde sur les décisions de justice rendues par les juges.

Le droit n'étant pas contenu dans des lois générales et impersonnelles, ce sont les juges qui, par les décisions qu'ils rendent, créent les règles de droit. Dans chaque décision rendue par les juridictions, les juges indiquent les motifs déterminants de la décision. Ces motifs constituent la *ratio decidendi* aussi appelée la « raison de la décision ». Le principe exposé dans celle-ci deviendra source de droit en ce que les juridictions inférieures devront se conformer à ce nouveau principe et ainsi respecter le principe du *stare decisis* aussi appelé « règle du précédent ».

Les différents principes dégagés par les *courts* sont dotés d'une autorité variable selon la place qu'elles occupent dans l'organisation juridictionnelle. Par exemple, dans le système juridique d'Angleterre et du Pays de Galles, la cour qui a la plus grande autorité est la UK Supreme Court (qui remplace la House of Lords depuis 2009), puis viennent la Court of Appeal et la High Court. Dans cette hiérarchie, chaque juge doit rendre une décision conforme au principe énoncé par celle de la juridiction supérieure.

Une partie peut faire appel de la décision si elle estime que cette dernière est contraire aux règles de droit. L'appel doit cependant être autorisé par la juridiction qui a rendu la décision. Cela implique que le juge admette que la décision qu'il a rendue puisse être erronée. De ce fait, peu de décisions font l'objet d'appels, environ une centaine par an.

Si dans les pays de *common law*, les normes écrites sont rares, elles existent néanmoins et prennent la forme de *statutes*, qui ont pour vocation de limiter ou de corriger la jurisprudence des juges.

La justice dans les pays du droit civil

Le système juridique de droit civil est un héritage romano-germanique, exporté dans les colonies européennes, qui est aujourd'hui utilisé dans toute l'Europe

et en Amérique du Sud, mais aussi en Russie et au Japon.

Le régime juridique civiliste est un système de droit dans lequel la loi, telle que votée par le Parlement, est obligatoire et s'impose dans les tribunaux. Les textes de lois sont rassemblés en des codes thématiques relatifs, par exemple, au travail, au commerce ou aux infractions pénales. La jurisprudence développée par les juges a moins d'importance que dans les pays de *common law* en ce que les juges n'ont pas l'obligation de rendre des décisions conformes à celles rendues antérieurement par leurs pairs.

L'article 5 du Code civil français précise ainsi depuis 1804 qu'il est « défendu aux juges de prononcer par voie de disposition générale et réglementaire sur les causes qui leur sont soumises ». Cette disposition légale interdit donc au juge civiliste de créer le droit par ses décisions contrairement au juge de *common law*.

De plus, pour motiver sa décision, le juge ne peut se borner à se référer à une décision antérieure intervenue dans une autre cause. Cela signifie que la jurisprudence n'est pas une norme obligatoire sur laquelle le juge peut se fonder exclusivement pour décider de l'issue d'un litige. Le juge civiliste doit donc travailler sur l'interprétation des textes de loi afin d'en dégager la solution du litige. Il peut s'aider de la jurisprudence en s'inspirant des décisions rendues par ses pairs dans des cas d'espèces proches. Cela permet de donner à la règle de droit une interprétation qui réponde le mieux au cas d'espèce. Enfin, le juge peut aussi être chargé d'interpréter les contrats conclus entre les parties ; en ce sens, il doit rechercher la commune intention des contractants, sans se limiter nécessairement aux termes de la convention.

Ainsi, si les juges n'ont pas juridiquement l'obligation de suivre les décisions rendues par les juridictions précédentes qui ne créent pas de droit, la jurisprudence de ces dernières bénéficie d'une autorité de fait ; notamment concernant les décisions rendues par la Cour de cassation ou le Conseil d'État, les deux juridictions suprêmes des ordres judiciaires et administratifs. En ce sens, les juges vont se conformer au raisonnement tenu par les juges hiérarchiquement supérieurs, au risque de voir leur décision cassée par ces derniers. Ainsi, le juge a moins de marge de manœuvre dans la création du droit et doit motiver ses décisions au travers des dispositions légales.

En somme, le droit civil est un droit écrit. Il édicte des normes générales et abstraites conçues par l'autorité législative. La jurisprudence vient affiner l'application de ces normes, par les décisions et jugements rendus par l'autorité judiciaire. Les magistrats disposent d'un pouvoir souverain d'appréciation, selon un raisonnement déductif (syllogisme), en se fondant sur la règle générale et en justifiant son application à un cas particulier. Autrement dit, il s'agit d'interpréter la règle générale par un va-et-vient avec le cas particulier. Parce qu'examiner les précédents n'est pas suffisant, cette approche est moins conciliable avec la justice algorithmisée. D'autant plus que les choix des juges résultent également d'une multitude de critères largement discrétionnaires et non nécessairement formalisables.

L'interprétation du juge étant par nature subjective, la décision qui en résulte peut varier en fonction de celui qui la rend, d'autant plus que les magistrats ne sont pas juridiquement liés par les décisions rendues par les autres juges. De ce fait, la systématisation des décisions de justice rendues par les juridictions appliquant une logique civiliste semble complexe. En effet, l'intelligence artificielle ayant des capacités limitées, il est aujourd'hui impossible de faire appliquer à une machine un raisonnement humain d'interprétation de textes de loi au regard de faits d'espèce. Le système de droit civiliste ne semble donc pas être le plus adéquat au développement de cette justice algorithmisée.

À l'inverse, le système de *common law* semble s'y prêter davantage en ce que les décisions rendues par les juridictions qui l'appliquent sont particulièrement exhaustives et se fondent sur des précédents ; ce qui rend le droit plus prévisible. De ce fait, il est plus simple de réduire des décisions de justice en un raisonnement permettant d'estimer la probabilité de l'issue d'un litige. Il s'agit de confronter, d'une part, une règle et ses applications antérieures (le précédent) et, d'autre part, des faits nouveaux. Cela expliquerait aussi pourquoi la justice prédictive s'est développée davantage dans les pays anglo-saxons que dans les pays de droit civil.

Cela ne veut pas dire pour autant que le système de droit civiliste est perméable à toute forme d'algorithmisation. En effet, les usages possibles de l'intelligence artificielle dans le système judiciaire ne doivent

pas nécessairement se traduire par le « remplacement » du raisonnement juridique ou être limités à la notion de « justice prédictive ». Il serait tout à fait envisageable, considérant l'état de l'art actuel, de programmer des logiciels capables d'identifier les arguments de faits et de droit, déterminants dans la motivation de décisions de justice et portant sur des espèces similaires. Puis, de les comparer avec les circonstances particulières du cas qui lui est présenté, afin de donner une liste des différentes solutions envisageables selon leur ordre de probabilité.

La justice algorithmisée en pratique

Case Law Analytics : l'évaluation chiffrée de réussite d'une affaire

En France, des startups comme Predictice ou encore Case Law Analytics offrent des outils prédictifs d'évaluation des chances de succès pour des procédures, par exemple de prestations compensatoires ou de pensions alimentaires. Il s'agit de litiges répétitifs plutôt que de contentieux complexes.

L'approche de la startup française Case Law Analytics consiste à entraîner un algorithme de classification pour prédire une distribution de probabilité sur les jugements possibles ; l'enjeu étant de prendre en compte les aléas humains et judiciaires inhérents à toute décision de justice. Cette approche de gestion du risque par l'IA et l'humain illustre parfaitement la nécessité de rassembler l'expertise des métiers de la justice avec celle des praticiens de l'IA pour co-construire le modèle[26].

Les premières phases de traitement de la donnée consistent :

– à segmenter par type de domaine juridique (droit de la famille, droit boursier, droit civil, etc.) ;

– à identifier la meilleure jurisprudence pour chaque cas de la base de données et à demander à des juristes spécialisés d'identifier des critères invariants sur lesquels le juge pourrait s'appuyer pour établir son jugement. Dans cette tâche, ils sont partiellement aidés par la machine, capable de détecter des schémas récurrents (*patterns*) dans les données ;

– générer une distribution de probabilité sur les jugements possibles grâce à un algorithme de *machine learning* de classification entraîné pour. Cette opération permet d'obtenir une pondération sur le spectre des jugements du plus favorable au plus défavorable. L'apprentissage de l'algorithme est ensuite validé à l'aide de données de test « fraîches » qui n'ont pas été vues lors de l'entraînement.

Un tel outil est susceptible de bénéficier à tous les acteurs du monde judiciaire.

Du côté des juges, il faciliterait leur travail en leur fournissant une systématisation des motivations envisageables, tout en les incitant à mieux détailler leur décision. Dans les cas où un juge souhaiterait ne pas suivre la solution la plus probable, il serait incité à détailler quelles ont été les circonstances particulières et déterminantes justifiant sa décision. L'explicabilité de la jurisprudence s'en trouverait ainsi augmentée, ce qui bénéficierait autant aux auxiliaires de justice qu'aux justiciables eux-mêmes. Du côté du justiciable, une telle algorithmisation constituerait un outil d'aide à la décision précieux car il rationaliserait l'aléa judiciaire[27]. Détenteur d'une telle information, le justiciable pourrait décider plus facilement de porter son affaire devant un juge ou préférer un des modes alternatifs de règlement des différends (MARD).

Enfin, pour l'avocat, ce type d'algorithme permettrait de répondre aux exigences de célérité des justiciables

26. Jean-Pierre Clavier (dir.), *L'Algorithmisation de la Justice*, Bruxelles, Larcier, 2020.
27. Bruno Dondero, *Justice prédictive : la fin de l'aléa judiciaire ?*, Paris, Dalloz, 2017.

et d'augmenter sa crédibilité en appuyant son conseil sur une statistique, s'il l'estime nécessaire. Notons qu'au Canada, les juges de la Cour supérieure de justice de l'Ontario[28] ont déjà refusé de retenir le poste dépenses de recherches juridiques de 900 dollars canadiens, car l'avocat n'avait pas utilisé les outils d'intelligence artificielle, alors même que ces derniers auraient permis de réduire significativement le temps de préparation du dossier.

Concernant l'évaluation de réussite d'un avocat pour une affaire déterminée, ou les chances de succès en fonction des juges qui siègent, de telles pratiques nécessiteraient de noter et/ou profiler les personnes concernées. À ce titre, la loi n°2019-222 du 22 mars 2019 de programmation et de réforme pour la justice interdit, en l'état, en son article 1.2.7. : « Le profilage des magistrats et des fonctionnaires du greffe sera également interdit afin de ne pas porter atteinte au bon fonctionnement de la justice. » L'article préliminaire du Code de la justice administrative confirme : « Les données d'identité des magistrats et des fonctionnaires de greffe ne peuvent faire l'objet d'une réutilisation ayant pour objet ou pour effet d'évaluer, d'analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou supposées. »

PredPol et la police algorithmisée

La police prédictive fait partie de la justice prédictive, comme étape qui la précède. Le logiciel PredPol (pour *Predictive Police*) ici présenté s'intéresse à deux applications concrètes, la suggestion aux policiers des zones de patrouilles quotidiennes et la liste des gens (avec un casier judiciaire) à approcher ou à suivre par la police.

En 2010, une équipe de l'université de Californie à Los Angeles dirigée par le professeur Jeffrey Brantingham cherche à établir des corrélations entre les crimes au cours du temps. En pratique, par analogie avec la détection des répliques sismiques, ils cherchent à modéliser et à prédire sur des temps relativement courts les futurs crimes. L'idée est simple *a priori* : dans les séismes et dans les crimes, il existerait une dépendance en temps et en espace selon la même dynamique et les mêmes effets d'occurrences. Les scientifiques de l'UCLA adaptent alors le modèle de prédiction sismique de Marsan[29], du nom du scientifique français David Marsan, à la récidive criminelle. Ce modèle est entraîné sur les crimes du passé, déclarés dans la ville considérée. Le type de crime, le lieu, l'heure, ainsi que le profil du criminel font partie des données d'apprentissage.

En 2012, le logiciel PredPol intégrant le modèle algorithmique entraîné est commercialisé *via* la société du même nom. La ville de Los Angeles est l'une des premières à utiliser cet outil au sein de ses bureaux de police. En pratique, PredPol estime tous les matins et pour les douze prochaines heures les lieux des futurs crimes, ainsi que les personnes potentiellement suspectes à approcher. La police peut ainsi prioriser les lieux de patrouilles quotidiennes et les personnes à surveiller. Même si des policiers affirment que le logiciel les aide au quotidien, d'autres pointent le manque d'efficacité de l'outil, pire, certains le considèrent comme discriminatoire envers certaines populations[30].

En effet, l'identification par le logiciel PredPol des lieux de crimes potentiels et du type d'individus jugé suspect peut être considérée comme discriminatoire. Tout d'abord, parce que l'algorithme a été entraîné sur les caractéristiques des crimes déclarés et non réellement commis. Autrement dit, l'algorithme risque de ne pointer que les mêmes zones ; les patrouilles policières désertant ainsi certains quartiers.

Concrètement, en extrapolant la fréquence des interpellations des individus pour anticiper leurs crimes et délits futurs, des catégories entières de populations pourraient être stigmatisées, comme les populations noires ou hispaniques. En d'autres termes, selon

28. Voir Anita Balakrishnan, « Judge says AI could have been used », *Law Times*, 3 décembre 2018.
29. David Marsan et Olivier Lengliné, « Extending Earthquakes' Reach Through Cascading », *Science*, n° 319, 22 février 2008, pp.1076-1079.
30. Scott Friedman et Jack Douglas, « LA's lessons for Dallas on Big Data Policing », NBCDFW, 28 février 2020.

l'algorithme, un ancien délinquant noir présentera de manière systématique un risque de récidive bien plus élevée qu'un ancien délinquant blanc. Il s'agit de discrimination raciale. Les travaux de Ramchand[31], par exemple, soulignent que certaines minorités sont plus susceptibles de se faire arrêter que d'autres.

Enfin, l'analogie faite avec la modélisation des répliques sismiques est relativement bancale. En effet, contrairement à la prédiction d'une récidive, la détection d'une réplique sismique n'engendrera jamais une autre secousse. Dans le cas des crimes, on observe un effet mécanique : l'algorithme se nourrit lui-même et gonfle ses propres prédictions. Ce qu'on appelle les *self-fulfilling predictions* ou prédictions auto-réalisatrices. Autrement dit, on observe une augmentation des arrestations dans les lieux suggérés par le logiciel, les lieux où le taux de crime est le plus élevé. Il apparaît ainsi que la stigmatisation de certains types de population les pousse à commettre un crime.

De manière générale, les algorithmes de police prédictive comme PredPol risquent de se concentrer sur les effets et les statistiques, avec les conséquences que l'on sait en termes d'allocation de fonds sécuritaires au détriment de l'identification des causes socio-économiques ou autres de la criminalité.

Il existe en Europe des systèmes de surveillance comparables. Financé par l'Union européenne, INDECT est un logiciel orienté sur les menaces terroristes et criminelles. Par une analyse combinée des données de la vidéo croisée avec celles extraites du web et des fichiers de police, il systématise l'investigation. Precobs en Allemagne ou encore Squeaky Dolphin[32], le programme de surveillance britannique du Government Communications Headquarters, surveillent en temps réel l'activité des réseaux sociaux, tels que YouTube, Facebook, ou encore Twitter.

COMPAS et l'évaluation algorithmisée de récidive

Le logiciel COMPAS, pour Correctional Offender Management Profiling for Alternative Sanctions, qui permet de prédire les risques de récidive d'un accusé ou d'un condamné, est sûrement l'exemple le plus connu dans la justice américaine. COMPAS est utilisé dans plusieurs juridictions du pays, dont celles des États de New York et de Californie. L'algorithme a été entraîné sur les récidives et les non-récidives du passé, et plus particulièrement sur le profil des personnes en question. Le logiciel analyse les réponses fournies par l'accusé ou le condamné à un questionnaire. Le logiciel formule ensuite une estimation de son risque de récidive. Les questions, relativement simples, portent, entre autres, sur le lieu de résidence de l'individu en question, sur ses antécédents judiciaires, mais aussi sur ceux de son entourage.

Ce type de modèle a apporté une certaine structure au jugement grâce à la découverte (supervisée ou non) de règles permettant d'établir une potentielle récidive. Ainsi, les travaux de Danner et de son équipe[33] prétendaient qu'on pouvait combiner ces outils avec l'expertise humaine afin de réduire certaines peines de prison d'individus qui ne présentaient pas de danger pour la société.

Malgré ces quelques succès, COMPAS présente de nombreuses failles et soulève plusieurs questions éthiques. En particulier sur les discriminations technologiques qu'il engendre en stigmatisant certaines populations comme la minorité noire[34] qui se voit accorder un score de risque systématiquement plus élevé. Les mécanismes de formation de biais algorithmiques au sein de COMPAS sont comparables à ceux des biais discriminatoires de PredPol envers les Noirs et les Hispaniques. Sans connaître le dessous du capot, on imagine facilement ses limites de

31. Rajeev Ramchand, Rosalie Liccardo Pacula et Martin Y. Iguchi, « Racial differences in marijuana-users' risk of arrest in the United States », *Drug Alcohol Depend*, vol. 84, n°3, 5 avril 2006, pp. 264-272.
32. « Squeaky Dolphin for sale: How surveillance companies are targeting social networks », *Privacy international*, 29 janvier 2014.
33. Mona J.E. Danner, Marie VanNostrand et Lisa M. Spruance, « Race and gender neutral pretrial risk assessment, release recommendations, and supervision », université Pretrial, novembre 2016.
34. Julia Angwin, Jeff Larson, Surya Mattu et Lauren Kirchner, « Machine Bias », ProPublica, 23 mai 2015.

fonctionnement liées entre autres aux types de données utilisées en apprentissage et à leur collecte. En effet, peu importe le type d'algorithme implicite utilisé – qu'il soit d'apprentissage statistique ou d'apprentissage sur réseaux neuronaux –, rien ne garantit que les données utilisées ne soient pas biaisées. De plus, même si ces données représentent correctement la population récidiviste américaine, elles ne peuvent pas devenir seules un maître-étalon de la récidive. Bien sûr, elles peuvent apporter un éclairage, mais certainement pas imposer un score *stricto sensu* sur lequel le juge se baserait pour trancher de manière unilatérale. Et bien que la couleur de peau ne soit pas demandée au sein du questionnaire et ne soit donc pas un des paramètres du modèle, on sait que par corrélation le lieu de résidence renseigne très souvent sur l'ethnicité du sujet. Ainsi, près de 70 % des habitants du quartier central Harlem de New York sont noirs, et plus de 60 % du quartier de Watts de Los Angeles hispaniques.

De même, contrairement à la société qui est dynamique, la *data* d'apprentissage n'est pas évolutive dans le temps ; cela induit un autre biais d'échantillonnage. Le logiciel ne prend, par exemple, pas en compte le changement de type de population dans une zone géographique donnée (vieillissement ou gentrification, par exemple), le changement de législation ou de jurisprudence, ou encore le contexte politique. Enfin, la question de la garantie de l'équité se pose. Pour valider et certifier un modèle, il faut évaluer si les décisions qu'il fournit sont équitables envers chaque individu ou chaque groupe d'individus. Pour ce faire, il faut traduire ce concept issu de la philosophie politique, en formalisme mathématique.

Parmi l'ensemble des formulations mathématiques disponibles[35], l'équité peut être définie mathématiquement au travers des trois grandes classes de mesures suivantes :

– l'anti-classification mesure l'aptitude de l'algorithme à ne pas utiliser, dans sa décision finale, des données d'entrée sensibles telles que la race, le genre et leurs dérivés. L'anti-classification résulte du phénomène sous-jacent (social, sociétal, économique, etc.) qui génère les données. Si l'utilisation de ce type de données est propice à créer des discriminations, il existe aussi des cas dans lesquelles l'étanchéité de l'algorithme aux caractéristiques intrinsèques d'un individu empêche une représentation exacte de la réalité. À titre d'exemple, pour des casiers judiciaires équivalents, des experts ont noté que les femmes sont moins susceptibles de récidiver que les hommes[36]. Ainsi, dans le dataset COMPAS, un algorithme n'utilisant pas le genre pour sa classification va systématiquement surestimer la probabilité de récidive chez une femme. En effet, sans le genre spécifié, les hommes et les femmes auront les mêmes risques de récidives par défaut, ce qui s'oppose à la réalité où il apparaît que les femmes récidivent beaucoup moins. Ceci a été prouvé dans le comté de Broward en Floride, où des femmes avec un risque COMPAS de 6/10 ont récidivé en proportion égale avec des hommes dont le score était de 4/10[37]. Constatant ce fait, la cour de justice du Wisconsin a autorisé l'emploi d'algorithme utilisant le genre comme critère pour ses règles de décisions, à condition que cette utilisation serve « l'institution ou le justiciable sans crainte de discrimination[38] ». Enfin, mesurer la sensibilité d'un algorithme à ce type de données garantit davantage d'équité et de compréhension du système qu'une stratégie où l'on enlèverait ces données, qui sont, de toute façon, prédictibles par le modèle, à partir d'autres caractéristiques corrélées ;

– la reproductibilité et la cohérence caractérisent la capacité des modèles entraînés sur un même jeu de données à fournir les mêmes résultats. Dans le cas de COMPAS, par exemple, plus de 40 % des résultats obtenus à partir de plusieurs modèles à performance comparable dépendaient du type de modèle. Ces résultats étaient donc différents d'un

35. Voir sur YouTube : «Tutorial: 21 fairness definitions and their politics par Arvind Narayanan », 2018.
36. Sharad Goel, Ravi Shroff, Jennifer L. Skeem et Christopher Slobogin, « The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment », 2019.
37. *Ibid.*
38. State v. Loomis – 2016 WI 68, 371 Wis. 2d 235, 881 N.W.2d 749.

modèle à l'autre[39]. La classification de plus de 40 % de personnes dépend du choix du modèle. Mesurer la cohérence du modèle finalement retenu permet de s'abstraire de ces effets néfastes pour l'égalité de traitement des individus ;

– la classification paritaire est une notion d'équité de groupe. Elle consiste à vérifier que chaque groupe défini comme sensible (une minorité, par exemple) soit traité de la même façon que la majorité. Ainsi, dans le dataset COMPAS, on observe un taux de faux positifs deux fois plus élevé pour la population afro-américaine dans le comté de Broward en Floride[40]. Cela étant dit, cette notion d'équité est discutable, car elle ne garantit pas qu'à l'intérieur d'un groupe on classe à tort plus d'individus comme susceptibles de récidiver que dans un autre groupe[41]. Afin de remédier à ce problème, les notions d'égalité des chances et d'égalité en opportunités ont été aussi considérées. On s'assure alors que les prédictions du modèle sont indépendantes des variables sensibles conditionnellement au vrai résultat ; c'est-à-dire les catégories (récidive ou non) auxquelles appartiennent effectivement les individus ;

– la calibration est, quant à elle, une notion d'équité individuelle. Des individus avec des scores COMPAS similaires doivent récidiver à des taux similaires, quels que soient leur origine ou le groupe ou les groupes auxquels ils appartiennent. Un modèle est dit calibré si une fraction p de l'ensemble des individus ayant reçu une probabilité de récidive p va effectivement récidiver. L'équité entre deux groupes voudrait que la calibration soit respectée à l'intérieur de ces deux groupes. En effet, supposons que le modèle ne soit pas calibré par rapport au groupe minoritaire, mais seulement par rapport à l'ensemble de la population comprenant la majorité et le groupe minoritaire. Dans le groupe minoritaire, le décalage entre la perception du modèle, qui classe davantage d'individus comme étant plus à risque, et la réalité est potentiellement plus important que dans le groupe majoritaire. Ainsi, intuitivement, les probabilités d'un modèle calibré sur tous les groupes prédéfinis ne contiennent pas d'information spécifique à un groupe[42].

Pour dépasser ces limites, il faudrait entraîner des modèles qui assurent des scores de risques corrélés, avec tous les critères qui définissent un individu dans sa spécificité, et non corrélés avec un groupe ou une appartenance sociale, ethnique, ou encore politique.

DataJust et le système d'évaluation automatique des indemnités judiciaires

Lorsqu'une personne a subi un dommage corporel et souhaite en obtenir l'indemnisation, conformément aux règles de la responsabilité civile, il revient aux juges d'évaluer le dommage subi et de décider du montant de l'indemnisation à attribuer. Pour l'évaluation du dommage, les magistrats ont recours à la nomenclature dite Dintilhac, qui liste les différents chefs de préjudices corporels. Cependant, la détermination du montant des indemnités est un processus complexe pour lequel les outils existants ne permettent pas d'assurer une homogénéité dans les réponses apportées par les avocats et les assureurs.

L'article 1269 du Code civil issu du projet de réforme de la responsabilité civile consacre le recours à la nomenclature Dintilhac[43] pour l'évaluation des préjudices. L'article 1271 précise qu'un « décret en

39. Charles T. Marx, Flavio du Pin Calmon et Berk Ustun « Predictive multiplicity in classification », ICML, 2020.
40. Jeff Larson, Surya Mattu, Lauren Kirchner et Julia Angwin « How we analyzed the compas recidivism algorithm », ProPublica, 23 mai 2016.
41. Brian Hu Zhang, Blake Lemoine et Margaret Mitchell, « Mitigating unwanted biases with adversarial learning », InProceedings of the 2018 AAAI/ACM Conference on AI.
42. A. Philip Dawid, « The well-calibrated Bayesian », *Journal of the American Statistical Association*, vol. 77, n° 379, septembre 1982 et Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg et Kilian Q. Weinberger, « On fairness and calibration », dans I. Guyon et al., *Advances in Neuronal Information Processing System*, New York, Curran Associates Publishers, 2017.
43. Cette nomenclature reprend l'ensemble des préjudices possibles en matière de dommages corporels tant pour vous (victime directe) que pour vos proches (victimes indirectes).

Conseil d'État fixe les postes de préjudices extrapatrimoniaux qui peuvent être évalués selon un référentiel indicatif d'indemnisation […]. À cette fin, une base de données rassemble […] les décisions définitives rendues par les cours d'appel en matière d'indemnisation du dommage corporel des victimes d'un accident de la circulation. »

La volonté de créer un outil d'aide à la détermination du montant de l'indemnisation du dommage corporel existait donc depuis environ deux ans. L'ancienne ministre de la Justice, Nicole Belloubet, par un décret du 27 mars 2020 publié le 29 mars 2020, a autorisé la mise en œuvre d'un traitement automatisé de données à caractère personnel, appelé DataJust, pour une durée de deux ans. Ce décret s'inscrit donc dans le cadre du projet de réforme de la responsabilité civile. L'objectif final de ce traitement de données est de mettre à disposition du public un référentiel indicatif d'indemnisation des préjudices corporels. Ce référentiel sera destiné à aider, d'une part, les parties et leurs avocats à estimer le montant des indemnités auxquelles elles peuvent prétendre en réparation des préjudices et, d'autre part, d'aider les magistrats à allouer aux victimes une indemnité juste.

Le décret permet la collecte automatique de données issues de décisions de justice et leur traitement par un algorithme. L'algorithme se nourrira des décisions rendues par les cours d'appel entre le 1er janvier 2017 et le 31 décembre 2019, et en particulier des données relatives aux montants alloués à la victime en fonction de la gravité des dommages corporels subis tels qu'évalués grâce à la nomenclature Dintilhac. En pratique, les décisions du Conseil d'État et de la Cour de cassation, conservées respectivement dans les bases de données Ariane et JuriCA, seront, elles aussi, analysées pour qu'en soient extraites les informations relatives aux indemnités accordées pour chaque type de préjudice.

Cependant, certaines critiques ont été formulées à l'encontre de ce traitement, principalement sur le risque de déjudiciarisation du règlement des litiges. En effet, si les victimes et les assureurs ont accès à un outil leur permettant d'estimer le montant des indemnités généralement accordées pour des dommages identiques à ceux subis au cas d'espèce, cela favoriserait le règlement amiable des litiges, mais cela pourrait également diminuer le recours à l'assistance des avocats.

Enfin, le syndicat de la magistrature a émis des inquiétudes relatives à la standardisation des décisions de justice. Si les juges restent libres dans leur appréciation, ils seront orientés dans leur travail par le référentiel. Cela pourrait conduire à l'homogénéisation de la jurisprudence au niveau national. Le risque étant que les spécificités de chaque situation ne soient pas prises en compte.

Cette préoccupation pose la question de l'adaptabilité de la norme juridique aux évolutions économiques et sociales. Comment découvrir les préjudices existants ou même en découvrir de nouveaux si la décision du juge est enfermée par le résultat supposé le plus probable ? Ce type d'incitation à suivre le résultat donné par l'algorithme comme étant le plus probable porte le nom d'« effet performatif ».

Dans le cas de DataJust, le risque de mettre à bas le libre-arbitre du juge reste relativement faible, car le décret vise plus à concevoir et à entraîner un outil d'aide à la décision qu'à mettre en œuvre une prise de décision automatisée. Par ailleurs, l'écueil de l'effet performatif n'est pas une fatalité. Il pourrait être contourné par la présentation d'une liste de solutions envisageables selon leur probabilité, plutôt que par la présentation de la solution la plus probable uniquement. L'explicabilité de l'algorithme, ainsi que la transparence des critères pris en compte par ce dernier contribueraient également à laisser au juge sa liberté de choix. En effet, cela lui permettrait de faciliter la comparaison entre les situations afférentes aux solutions présentées par l'algorithme et celle dont il doit traiter.

Reconnaissance faciale et prédiction de la criminalité

Beaucoup plus exploratoire, un groupe de recherche de l'université de Harrisburg a développé en mai 2020 une IA de reconnaissance prétendument capable de prédire si quelqu'un va être un criminel, sans biais racial, seulement en se basant sur l'image du visage de la personne. Une pétition « Abolish the Tech to

prison pipeline »[44] a été mise en ligne pour s'opposer à la parution de l'article dans la revue *Springer*. Plusieurs scientifiques et experts du domaine tels que Pawel Drozdowski[45], Timnit Gebru, Margaret Mitchell ou Deborah Raji[46/47] dénoncent ainsi les travaux hasardeux de l'IA appliquée à la procédure judiciaire ou la criminalité et ceux qui utilisent des jeux de données biométriques ou issus de la biologie. La plupart des algorithmes d'IA ne font que trouver des corrélations entre les données d'entrées dont ils se servent pour s'entraîner et les résultats de la tâche de sortie à effectuer. Or, il est bien établi que l'on peut trouver facilement un grand nombre de corrélations absurdes[48].

Les nouvelles technologies de prédiction devront respecter les principes du droit pénal. Or, elles sont conçues sur des critères de dangerosité des individus – c'est-à-dire leurs actions possibles dans le futur – et non sur l'évidence de culpabilité – exigeant la preuve de faits commis. Le risque est de voir sanctionné demain l'écart à une prétendue norme sociale ou économique définie par un code informatique, sans rapport avec la norme juridique. On ne punirait plus un sujet pour ses actions, mais pour son profil dans une situation donnée. Quelle sera alors la référence ? La professeure de droit Mireille Delmas-Marty nous prévient : « À terme, c'est la disparition entre armée et police, ennemi et criminel, et, finalement, la confusion entre guerre et paix, qui sont ainsi programmées[49]. » La question de la place du droit pénal, des garanties procédurales et, partant, de l'État de droit devront être posées avec insistance.

Cette question devient aujourd'hui fondamentale tant elle tient à l'avenir de la démocratie. Que se passera-t-il lorsque les données désigneront des criminels avant même qu'ils n'aient commis leurs crimes ? Que restera-t-il de la présomption d'innocence pour celui qui présente les caractéristiques d'un multirécidiviste ? Il s'agirait alors de glisser insensiblement du commencement d'exécution à l'acte préparatoire, puis de l'acte préparatoire à la potentialité de commettre un crime. Indéniablement, cette piste ne peut être envisagée qu'avec intérêt par les organismes publics ou privés en charge de la sécurité, que ce soit pour des raisons de sécurité publique ou d'intérêt commercial. La plus neuve des technologies, par un extraordinaire raccourci, rejoindrait alors la plus vieille des criminologies. Les systèmes numériques prédictifs permettraient d'identifier le criminel en puissance. Tout comme les caractéristiques morphologiques devaient permettre d'identifier le criminel-né, selon le père de la criminologie italienne, Cesare Lombroso[50].

La présomption d'innocence est un principe fondamental de notre État de droit et l'une des premières caractéristiques de nos démocraties libérales. Elle permet de savoir les faits qui nous sont reprochés, de participer aux audiences et d'être entendus. Ces principes devront être respectés par les concepteurs et les utilisateurs de ces nouveaux outils.

44. « Help Fight the #TechToPrisonPipeline », Coalition for Critical Technology, 23 juin 2020.
45. Pawel Drozdowski et *al.*, « Demographic bias in biometrics: A survey on an emerging challenge », *IEEE Transactions on Technology and Society*, 2020.
46. Inioluwa Deborah Raji et *al.*, « Saving face: Investigating the ethical concerns of facial recognition auditing », *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 2020.
47. Inioluwa Deborah Raji et *al.*, « Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing », *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, janvier 2020.
48. Comme mis en évidence sur le site www.tylervigen.com/spurious-correlations.
49. Mireille Delmas-Marty, *Aux quatre vents du monde*, Paris, Seuil, 2016.
50. Cesare Lombroso, *L'Homme criminel*, 1876.

Précautions dans la justice algorithmisée

Les biais algorithmiques, sources de discrimination

Depuis la récente démocratisation de l'IA, les données sur lesquelles les modèles sont entraînés ont souvent été comparées au nouveau pétrole à forer. Mais la donnée n'est pas l'ultime objectif, et ne suffit pas à faire de bons modèles. La justice algorithmisée montre qu'il est nécessaire d'être précautionneux : la plupart de nos données sont davantage le reflet, même partiel ou déformé, des citoyens, des consommateurs ou simplement de l'humain. Elles doivent être manipulées avec discernement et un sens critique affûté.

Les cas présentés dans ce rapport, qui incarnent les risques inhérents de la prédiction algorithmique, sont quelques exemples de biais algorithmiques provenant d'un mauvais échantillonnage des données d'apprentissage, d'un manque de diversité et de représentativité de ces *data*, ou encore d'une systématisation de la réponse algorithmique à partir d'une certaine statistique à un temps donné.

Les conséquences sont le traitement inégal, stigmatisant et injuste des individus et des situations. Ces biais proviennent, entre autres, de ceux des concepteurs, mais aussi de ceux issus de la société elle-même. Ils tendent à réduire les êtres et les choses à leur description la plus simpliste, souvent approximative et trop souvent fausse. Les concepteurs et les utilisateurs (policiers, juges, ou encore avocats) ont en cela les mêmes responsabilités vis-à-vis de ces algorithmes.

Dans la justice algorithmisée, le risque premier, et sûrement le plus dommageable, serait de ne prendre en considération ni les causes ni les intentions, en gommant l'existence même des personnes, pour ne gouverner qu'à partir d'une expression statistique de la réalité. Comme l'analyse Antoinette Rouvroy, on passerait alors d'une logique de prévention de type probabiliste à une logique de préemption, qui modélise l'environnement, afin qu'un comportement à risque ne puisse survenir[51]. Comme de mauvais payeurs qui seraient exclus du champ du crédit, ou encore comme des individus dangereux refoulés d'un stade ou d'un concert.

Il est fondamental qu'un algorithme ne reflète pas les préjugés de ses concepteurs dans l'inventaire des critères utilisés pour déterminer les menaces. À titre d'illustration, peu importe le niveau de complexité future des algorithmes, il devra toujours exister des raisons objectives avant une fouille policière. Le respect du principe selon lequel toute personne est réputée innocente tant que sa culpabilité n'a pas été légalement établie devra absolument être défendu à l'heure de la justice algorithmisée.

Perte d'explicabilité et d'interprétabilité, le risque d'une justice opaque

Dans les années 1820, le sociologue français Auguste Comte expliquait déjà dans son œuvre *Opuscules de*

51. Antoinette Rouvroy et Thomas Berns, « Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation ? », *Réseaux*, vol.1, n°177, 2013.

philosophie sociale[52] que les données ne définissent pas la société. Il critiquait fermement ceux qui prétendaient utiliser les probabilités pour rendre compte de la complexité des comportements humains.

Une donnée s'inscrit systématiquement dans un contexte particulier, caractérisée par des conditions de création et de collecte qui lui sont propres. En cela, l'absence de contextualisation des données et de leurs collectes est un frein à la transparence des mécanismes de fonctionnement des algorithmes, dans le domaine de la justice. Cette contextualisation inscrit également les choix plus ou moins arbitraires, parfois biaisés des humains en charge de la conception et du développement. Dans ce souci et cette recherche perpétuelle de transparence des pratiques judiciaires, il y a alors un impératif de reconstitution des processus historiques, sociaux, mais aussi économiques qui ont contribué à la création des ensembles de données finalement fournies à un algorithme.

Par exemple, les bases de données regroupant les statistiques pénales (comme Cassiopée et le Système d'information décisionnel[53]) sur les natures des affaires traitées par les juridictions, leurs auteurs, les victimes, mais aussi l'issue des procédures, n'indiquent pas le contenu des décisions judiciaires. Une analyse fine de ces décisions permettrait d'obtenir cette information. Mais cela supposerait un échantillon large de décisions et donc de données importantes et représentatives pour pouvoir souligner au sein de leurs contenus l'ensemble des *metadata* déterminantes.

Le processus de décision au sein et à partir de la suggestion algorithmique est également fondamental dans la mise en évidence d'une quelconque discrimination ou de l'existence d'un biais algorithmique. Les régulations sur la protection des données, telle que le RGPD (Règlement général sur la protection des données) en Europe, ou le CCPA (California Consumer Privacy Act) en Californie, ont réaffirmé le droit élémentaire des citoyens à recevoir un traitement équitable (appelé aussi droit à la « non-discrimination » dans le CCPA) lorsque l'algorithme traite leurs données à caractère personnel. En pratique, la transparence sur le processus de décision est une conséquence de multiples étapes dans la chaîne de construction d'un modèle d'IA. Elle dépend du type de modèle algorithmique utilisé, mais aussi et surtout de l'ensemble des étapes qui soutiennent la construction du modèle finalement entraîné. Parmi ces étapes, se trouvent la collecte des données, le prétraitement et la labellisation de ces données, l'implémentation du modèle, les tests avant déploiement, ou encore les procédés de surveillance[54] de l'algorithme une fois en usage qui inclut le risque de biais ou la performance calculatoire. Cette transparence sur le processus de décision dépend également du type d'explication recherchée de l'algorithme.

L'explication locale fournit une explication du comportement algorithmique pour une donnée en particulier. Par exemple, pour un algorithme d'estimation du risque de récidive, on s'attache à comprendre l'influence de paramètres d'entrée décrivant un profil de personne en particulier, sur la réponse algorithmique. L'explication globale livre la logique générale de l'algorithmique à produire une réponse, et donc une suggestion de décision. Pour un algorithme de mesure du risque de récidive, on s'assure que le genre ou la localisation géographique n'est pas un critère décisif pour le modèle en général. Enfin, les explications contrefactuelles, cas particulier d'explications locales, cherchent à comprendre la logique de l'algorithme dans des cas imaginés. Pour cela, elles tentent de répondre à des questions de l'ordre du « et si…? », qui impliquent la variation d'une ou plusieurs valeurs d'entrée.

Cette méthode permet de mettre en évidence les inconsistances dans la catégorisation implicite des individus et des situations en dépit de leurs similarités, et donc de risques de biais. Il existe d'autres méthodes, comme celles utilisant des modèles simples de type proxy pour expliquer localement la logique

52. Auguste Comte, *Opuscules de philosophie sociale. 1819-1828*, Ernest Leroux Éditeur, 1883.
53. « Les indicateurs statistiques pénaux trimestriels », ministère de la Justice, 29 avril 2021.
54. On parle aussi de *monitoring* en anglais.

algorithmique. En général, la plupart des méthodes d'explicabilité souffrent d'un risque de robustesse[55].

Il est évident que l'explicabilité des algorithmes dans le domaine de la justice est un enjeu, pour ne pas dire une obligation. L'enjeu est de garantir une certaine transparence des pratiques judiciaires pour les acteurs de la loi (police, juge, ou avocats), mais aussi, et surtout, pour l'accusé, la victime et le citoyen de manière générale. Autrement dit, en comprenant les mécanismes algorithmiques qui co-gouvernent son futur, à l'instar de sa compréhension de la loi, la justice ne peut pas voir disparaître le droit pour un accusé de se défendre.

55. Stratis Tsirtsis et Manuel Gomez-Rodriguez, « Manuel. Decisions, Counterfactual Explanations and Strategic Behavior », NeurIPS, 2020 ; Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru ErhanBeen et Kim Kindermans, « The (un)reliability of saliency methods » dans *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019.

Quelques recommandations

Le choix du modèle algorithmique

En AI, comme présenté au début de ce rapport, les modèles algorithmiques diffèrent par leur niveau de complexité, mais aussi d'explicabilité. Pour assurer un haut niveau d'explicabilité et de compréhension du modèle, dans l'objectif d'une transparence de la justice algorithmisée et donc de la justice, le choix du modèle est une décision stratégique. Des modèles linéaires ou des arbres de décisions avec des conditions entièrement explicites ou obtenues par apprentissage (implicites) vont être plus « explicables » que des modèles dont les paramètres, parfois au nombre de plusieurs milliards, se disposent dans des réseaux à plusieurs couches, résultants des méthodes dites d'« apprentissage profond » (*Deep Learning*).

Ces derniers modèles sont souvent plus performants dans l'exactitude et la précision de la réponse, mais ils requièrent également, pour être entraînés, l'utilisation d'infrastructures plus puissantes et optimisées pour le calcul matriciel. Pour tenter de limiter les compromis entre explicabilité et performance algorithmique, contraints par la technicité du matériel informatique (puissance des microprocesseurs, par exemple), de nombreux travaux de recherche se penchent sur le domaine de l'explicabilité (ou XAI). Cela constitue un enjeu majeur pour le déploiement et l'adoption de ces technologies et de ces algorithmes dans tous les secteurs régulés, dont la justice fait partie.

L'usage de modèles hybrides, contenant à la fois une composante algorithmique explicite (possiblement mathématisée) et une composante algorithmique implicite (issue d'un entraînement), permet d'augmenter le niveau d'explicabilité.

Des bonnes pratiques de développement

Comme précisé précédemment, la transparence du processus décisionnel algorithmisé contient également l'ensemble des étapes de construction, de test et de déploiement de l'algorithme en question. En cela, les pratiques de développement, incluant autant les discussions sur l'idée même de cet algorithme, sa programmation informatique, sa validation que sur son usage et sa compréhension par les utilisateurs, doivent être bien désignées pour assurer son bon fonctionnement et l'absence de risques de biais, par exemple. Cela passe par des automatismes dans l'écriture du code informatique, la manière de tester la *data* et le code, mais aussi dans la façon de relire le code de ses confrères au sein d'une équipe (on parle de *reviewing* de code). Les discussions préconception et post-développement peuvent également être orchestrées en incluant des profils non techniques, comme des personnes issues du métier. Dans le cas de la justice algorithmisée, des acteurs de la justice, comme les juges, les avocats, ou encore les policiers, seraient consultés. On parle alors de co-conception pour intégrer les bons réflexes d'une pratique judiciaire éthique, responsable et juste.

L'humain dans la boucle algorithmique

La justice est et restera toujours humaine et humaniste. En cela, l'idée du juriste remplacé par une machine est de l'ordre de l'imaginaire ; elle est à exclure. Au contraire, la justice algorithmisée bien pensée, responsable et éthique permettra de mieux comprendre les affaires dans les faits, le contexte mais aussi les

hommes et les femmes qui en font partie. D'un point de vue technologique, il sera difficile, voire impossible, de reproduire avec exactitude et exhaustivité le métier de juriste. Même entraîné au mieux à exécuter une tâche, un algorithme manquera toujours cruellement de sens commun propre à l'homme. Aussi, d'un point de vue éthique et déontologique, il y a le besoin de conserver un certain jugement humain dans la considération et l'appréciation d'affaires. Écrit autrement, l'humain doit rester dans la boucle, même lorsque des algorithmes sont utilisés pour éclairer une décision.

Insuffler une culture du numérique, former et hybrider les connaissances

Une innovation responsable et constructive ne pourra se développer dans notre pays qu'à condition de créer une culture du numérique. De nouvelles disciplines, qui se situent à l'intersection des sciences dures et des sciences humaines, doivent émerger : l'histoire des technologies, l'éthique des données, la sociologie, ou encore la robotique.

À l'heure des algorithmes, cela concerne également la justice. À cet égard, une hybridation des connaissances est nécessaire. Le codeur de demain devra intégrer des questions éminemment éthiques ; tout comme le juriste ne pourra pas être ignorant des statistiques et du fonctionnement des systèmes dits d'intelligence artificielle.

Afin d'explorer ces nouveaux champs d'étude, nous devons, à l'image de la présente étude, prôner la transdisciplinarité. En effet, nous ne pouvons pas nous contenter de raisonner par silo. L'accélération du temps induite par les technologies appelle à une confrontation des domaines d'étude.

Des outils pour évaluer les algorithmes

Le travail d'évaluation des algorithmes est un travail scientifique et technique, mais aussi humain. En cela, une collaboration étroite et transparente entre experts techniques, concepteurs et utilisateurs et/ou personnes du métier d'application (ici la justice) est fondamentale. Ainsi, au-delà du simple développement des modèles d'intelligence artificielle optimisés selon certaines métriques de précision ou d'équité prédéfinies, lorsque ces systèmes sont utilisés en pratique dans la sphère publique, le mécanisme de décision est aussi affecté par l'interaction entre les humains (ou l'opérateur qui utilise le modèle) et la machine. L'utilisateur du modèle peut ainsi vouloir optimiser selon des contraintes plus larges comme une orientation de politique pénale et la décision du modèle ne représente qu'une composante servant à la décision finale. Dès lors, le risque algorithmique doit prendre en compte une chaîne complexe de décisions (et non pas uniquement un modèle isolé) pour être mesuré avec précision.

Ainsi, une série d'outils collaboratifs associant les différentes parties prenantes pour mettre l'humain au centre du processus algorithmique afin de bâtir des outils de confiance sont développés :

– l'équipe DataSF de la ville de San Francisco associée aux universités de Harvard et de Johns Hopkins, ainsi que GovEx ont créé un questionnaire[56] suffisamment général pour être utilisable dans la plupart des projets algorithmiques. Il consiste en deux étapes : une auto-évaluation du risque algorithmique (identification des personnes impactées, le type de données utilisées, ou l'auditabilité du modèle) ; puis la résolution des problèmes potentiels grâce à l'analyse des réponses (gouvernance, discussion entre les parties prenantes ou tests automatiques systématisés par exemple) ;

56. http://ethicstoolkit.ai/.

– le département de droit de l'université de Stanford a développé un outil d'aide à l'évaluation des technologies émergentes dans la police[57]. Cet outil peut à la fois être utilisé pour sensibiliser des équipes aux précautions à prendre et aux risques à considérer lors de l'emploi d'outil algorithmique dans la police, ainsi que pour établir des bonnes pratiques dans la collecte, l'utilisation et la gouvernance des données de police ;
– l'Open Data Institute propose un outil[58] pour les développeurs de solutions recourant à l'algorithmique, pour le secteur public. On y trouve des études de cas spécifiques ou des définitions de cadres éthiques pour le passage à l'échelle de projets ;
– l'Union américaine pour les libertés civiles propose aux citoyens un guide[59] pour les aider à déterminer si une décision provient d'un modèle algorithmique, explorer l'impact de certaines technologies, notamment de surveillance, ou poser des questions importantes aux décideurs sur ces sujets.

57. Voir « Emerging Police Technology: A Policy Toolkit », Stanford Law School.
58. https://theodi.org/service/tools-resources/data-and-public-services-toolkit/.
59. www.aclu-wa.org/AEKit.

Conclusion

Le sujet de la justice algorithmisée est inégalement connu, compris et appréhendé. Une crainte affichée et partagée par les auxiliaires de justice est que l'utilisation d'outils automatisés contribuerait à la déshumanisation du processus judiciaire. En creux, la question est donc celle d'étudier les conditions d'acceptabilité qui permettraient à de nouvelles familles d'algorithmes d'être à terme légitimement utilisées au service du droit, des activités juridictionnelles, de la justice et des justiciables.

Car si d'une certaine manière la justice a toujours été en partie algorithmique au sens où elle émane d'un processus qui suit des étapes pour parvenir à une décision, elle a désormais vocation à être algorithmisée par des processus numériques et informatiques par le truchement d'outils d'aide à la décisions basés sur le traitement d'un volume important de données.

Les répercussions sont multiples. L'analyse d'une quantité exponentielle de données par des algorithmes peut concurrencer les sources de droit traditionnelles, à côté des lois, de la jurisprudence ou encore des contrats. Les juges, les avocats et les juristes vont recourir de manière croissante à des processus algorithmiques pour motiver leurs décisions, légitimer les arguments, justifier leurs positions. Ainsi, les parties au procès devraient à terme préciser les éventuels outils d'aide à la décision auxquels ils recourent. Pour qu'un justiciable reste en mesure de contester l'usage d'un outil algorithmique plutôt qu'un autre, tout comme les résultats en émanant (estimation, chance de probabilité, etc.), et ce afin d'éviter toute inégalité des armes ou de traitement.

Par ailleurs, le recours aux algorithmes incite à une approche inductive, fondée davantage sur la corrélation, plutôt qu'à une approche déductive, fondée sur la causalité. Dans ce contexte, le raisonnement par syllogisme – avec la majeure constituant la règle de droit, la mineure son application au cas d'espèce et la conclusion – risque d'être fortement concurrencé. La justice algorithmisée viendrait alors bouleverser le raisonnement jusqu'alors connu et qui servait de motivation aux décisions judiciaires.

Dans ce contexte, se pose la question de savoir comment ces algorithmes pourront prendre en compte des considérations sociales si importantes de cette matière mouvante et humaine qu'est le droit, par exemple en matière pénale lorsque sont appréhendés les éléments de personnalités d'un accusé. L'ensemble du corps judiciaire sera-t-il suffisamment formé, détiendra-t-il la culture numérique nécessaire pour raisonner avec et contre, remettre en question, contester les résultats proposés par des outils algorithmiques ? Autrement dit, dans quelle mesure serons-nous collectivement capables d'interpréter, de nuancer ces nouveaux *outputs* ? Autant de questions auxquelles il faudra apporter une réponse collective, en co-construisant les modalités de déploiement. Autant de réponses qu'il conviendra de penser afin d'éviter que la justice algorithmisée nous impose ces calculs, ou une logique qui nous échapperait, selon des mécanismes qui pourraient nous demeurer invisibles, incompréhensibles ou inintelligibles.

L'humain reste indispensable au bon fonctionnement de la justice. L'intelligence émotionnelle, la capacité à apprécier des situations inédites, ou encore celle à écouter et à débattre – au fondement du principe du contradictoire – sont des apports essentiels à l'acte de juger. La question qui se pose donc n'est pas celle du remplacement de l'homme par l'algorithme, mais plutôt celle de l'outillage des différents acteurs de la scène juridique et judiciaire.

Afin de placer l'intelligence artificielle sur le terrain de l'amélioration de la justice, il convient de se demander quelles solutions concrètes peuvent être mises en place pour équilibrer les rapports algorithme/juge, algorithme/justiciable et algorithme/avocat. Pour les juges, il conviendra notamment d'apprendre, entre autonomisation et automatisation, à

composer avec ces nouveaux outils d'aide à la décision, lesquels ne devraient réduire leur autonomie décisionnelle ou leur capacité à présenter la solution d'un litige. Pour le justiciable, l'algorithme devrait l'éclairer quant à l'opportunité de présenter son affaire devant un juge et dans l'effectivité de son droit d'accès à la justice. À cet égard, il reviendra probablement à l'avocat d'effectuer un travail de pédagogie auprès de son client afin de lui exposer toutes les options possibles et de lui expliquer qu'au-delà de leur chance de succès, chacune reste envisageable. Enfin, pour l'avocat, les algorithmes pourraient permettre de réduire les temps de recherche et de conforter sa légitimité, notamment pour les conseils concernant le choix de mener une action en justice ou non.

L'algorithmique n'est pas une fin, mais un moyen, un outil qu'il faut manipuler avec certaines connaissances et compétences pour garantir sa bonne compréhension et son utilisation. La justice, comme toutes les autres disciplines, profite, à raison, de l'émergence d'outils et de méthodes algorithmiques toujours plus efficaces pour améliorer sa propre efficacité, sa productivité et sa précision. La justice, du fait de l'implication d'êtres humains et de leurs histoires, mais aussi du fait de son influence sur la société, est un domaine d'application éminemment délicat.

L'enjeu sous-jacent à l'ensemble de ces interrogations est de déterminer si une justice algorithmisée sera ou non synonyme de progrès. La réponse à cette question dépendra en particulier du fait de savoir si la justice algorithmisée permettra d'apporter une meilleure administration de la justice, en renforçant la confiance portée par les citoyens dans les décisions rendues. Cela dépendra notamment de son accessibilité, de sa capacité à rendre des décisions plus rapides, de manière plus équitable, tout en garantissant les droits fondamentaux.

Table des matières

Réalisation : REFLETS GRAPHICS
Imprimé en France par A.Trois
JUIN 2021

Collection dirigée par Gilles Finchelstein et Laurent Cohen

12, cité Malesherbes - 75009 Paris

www.jean-jaures.org

Derniers rapports et études :

06_2020 : Défendre les droits des personnes intersexes :
pour une évolution ambitieuse du droit et des pratiques
Flora Bolter, Anne-Lise Savart

07_2020 : La rémunération du travail politique
sous la direction d'Éric Kerrouche et Rémy Le Saout

08_2020 : Construire la résilience territoriale pour anticiper les chocs à venir
Coordination « bouclier anti-Covid » des maires franciliens (COMIF)

08_2020 : Repenser notre fiscalité. Manifeste pour une imposition plus simple et plus équitable
Brice Gaillard

11_2020 : N'est pas métropole qui veut, ou le trompe-l'œil lyonnais
Vincent Aubelle

11_2020 : Repenser nos sociétés à l'aune des Objectifs de développement durable
sous la direction de Jennifer De Temmerman et Alain Dubois

03_2021 : La Protection salariale garantie
Amin Mbarki, Samuel Toubiana, Anthony Paulin

03_2021 : La raison d'être des entreprises : deux ans après, premier bilan

03_2021 : Travailler à l'âge du numérique : l'an II des coopératives !
Jérôme Giusti, Thomas Thévenoud

05_2021 : Élections européennes et Covid-19 : quelle visibilité de l'Union européenne
dans les journaux télévisés ?
Fanny Hervo, Théo Verdier

05_2021 : Signaler la haine pour mieux la combattre. Les LGBTphobies au prisme de
l'application FLAG!
Flora Bolter, Denis Quinqueton, Johan Cavirot

fondationjeanjaures
@j_jaures
fondation-jean-jaures
www.youtube.com/c/FondationJeanJaures

www.jean-jaures.org

Fondation Jean Jaurès ÉDITIONS