Algorithmes et droit pénal : quel avenir ?

Par **Elise BERLINSKI**
ESCP
Imane BELLO
Avocate à la Cour
Et **Arthur GAUDRON**
Chercheur au Centre de Robotique

« Les algorithmes brevetés inondent le système de justice criminelle. Les systèmes de *machine learning* déploient les agents de police dans les quartiers prétendus "chauds". Les laboratoires de police scientifique utilisent des logiciels probabilistes pour analyser les indices médico-légaux. Les juges utilisent des "instruments d'évaluation du risque" automatisés pour décider qui peut sortir sous caution, voire même quelle peine imposer » (Wexler, 2017).

Les algorithmes d'intelligence artificielle (IA), plus spécifiquement de *machine learning* (ML), sont à présent une réalité dans le droit pénal états-unien et ne sauraient tarder à le devenir en France, comme en témoignent par exemple les discussions sur leur application dans ce domaine du rapport de la CNIL (Demiaux & Si Abdallah, 2017). Dans cet article, nous nous intéressons aux implications de l'utilisation du ML dans le cadre du droit pénal français, et la manière dont celui-ci pourrait le modifier.

L'application d'algorithmes de ML dans le contexte du droit pénal vise à la création d'indices dans l'optique d'assister à la détermination du déroulement effectif d'un acte répréhensible, ou à l'orientation d'une enquête. L'utilisation de tels algorithmes demande de se poser des questions sur les données qui les alimentent, d'une part, et les calculs menés, d'autre part. Ces données proviennent de plus en plus fréquemment de réseaux sociaux, comme en témoigne le discours de Lord Gross (2018). Elles sont souvent comportementales, construites par désagrégation et réagrégation, et visent à reconstituer un sujet, généralement traité comme représentatif d'un sujet réel, donc capable d'en dévoiler l'essence (Chamayou, 2015), alors même que des études montrent qu'il n'existe pas de correspondance évidente entre ce sujet numérique construit et un sujet référent réel (Goriunova, 2019). Il nous semble donc urgent de comprendre le prisme de médiation (Hansen & Flyverbom, 2015) par lequel ces technologies représentent les personnes et leurs effets.

Après une rapide présentation du cadre juridique pénal en France, nous décrivons les méthodes de construction du sujet numérique, avant d'en tirer des conclusions sur le phénomène observé.

Les preuves et raisonnements juridiques en droit pénal

En droit pénal, une infraction est constituée par la réunion des éléments suivants : un élément légal (la répression par un texte de loi), un élément matériel (un comportement en lien de causalité avec le résultat entraîné) et un élément moral (la participation libre à un fait dont on connaît le caractère illégal). Les infractions peuvent être matérielles (le dommage est réalisé) ou immatérielles (l'accomplissement de l'acte incriminé seul suffit pour être qualifié d'infraction, c'est par exemple le cas de la fabrication de fausse monnaie). Par ailleurs, une tentative de crime est elle aussi répréhensible dès l'instant qu'« elle n'a été suspendue ou n'a manqué son effet qu'en raison de circonstances indépendantes de la volonté de son auteur » (code pénal - article 121-5, non daté). Il est donc envisageable que les officiers de police utilisent des algorithmes de ML de

manière à déterminer la probabilité qu'un crime ou délit ait lieu puis à déclencher une enquête en vue d'en déterminer son existence effective.

A titre d'exemple, l'administration fiscale (tant la Direction générale des finances publiques que la Direction générale des douanes et droits indirects) utilise des algorithmes de ML en vue de détecter des incohérences susceptibles de constituer des schémas de fraudes ou des signaux faibles de potentielles fraudes (Bello & Daoud, 2020).

Afin d'entrer en éventuelle condamnation, le juge doit analyser la caractérisation de l'infraction qui se situe dans le passé et se matérialise par un comportement. Cette analyse s'appuie sur des preuves dont le régime en droit pénal est dit « libre » : aux termes de l'article 427 du code pénal, « Hors les cas où la loi en dispose autrement, les infractions peuvent être établies par tout mode de preuve et le juge décide d'après son intime conviction. Le juge ne peut fonder sa décision que sur des preuves qui lui sont apportées au cours des débats et contradictoirement discutées devant lui. » Ainsi, le résultat de modélisations statistiques, de systèmes d'exploration de données algorithmiques, s'il ne peut pas, à proprement parler, servir comme élément de preuve de la réalisation d'un acte, peut toutefois influencer l'appréciation souveraine du juge.

Ainsi, bien qu'il n'existe pas de disposition légale réprimant un acte qui ne soit pas réalisé (de manière matérielle, immatérielle, ou sous forme de tentative), nous observons que les résultats des algorithmes peuvent influencer le juge (et sa perception du réel), d'une part, et, d'autre part, orienter l'enquête. Dans le premier cas, même si les résultats algorithmiques ne pourront être utilisés sans preuve tangible de la réalisation de l'acte, il est possible d'imaginer qu'ils puissent orienter la détermination de la peine et appuyer la conviction du juge. En effet, selon le principe d'individualisation des peines, la personnalité et la singularité d'une personne sont prises en compte dans le prononcé de sa peine. Dans le second cas, l'algorithme orientera le choix de réaliser une enquête et la façon dont celle-ci sera menée. Dans tous les cas, il nous semble nécessaire de comprendre le prisme pour penser le phénomène auquel nous faisons face.

Création du sujet numérique

Deux blocs constituent ce sujet : les données et les calculs algorithmiques qui visent à déterminer la probabilité qu'un individu ait commis une infraction ou non.

Les données décrivant le sujet auquel on s'intéresse peuvent provenir, d'une part, du domaine juridique (l'ensemble des éléments enregistrés liés à une affaire, pouvant s'étendre des déclarations à la police jusqu'au jugement), d'autre part, du domaine public/privé (collectées par un ensemble de capteurs numériques, dont les réseaux sociaux).

Ce profil singulier doit par ailleurs être « comparé » (par un modèle) à des bases de données d'observations, ou données d'entraînement, qui informent, d'une part, des infractions commises par une personne, d'autre part de leurs données comportementales collectées par divers capteurs numériques. À l'heure actuelle, les réseaux sociaux représentent de bonnes bases de données, puisqu'ils placent des capteurs sur l'ensemble du *web* de manière à constituer des profils comportementaux, qu'ils monétisent ensuite à des organisations qui veulent justement évaluer des profils de risque (crédit, assurance…) (Arvidsson, 2016 ; Fourcade & Healy, 2013).

L'algorithme de calcul n'est pas déterminé par une théorie (sociologique par exemple), mais se forme en apprenant à partir de la donnée et d'inférences statistiques. Nous nous concentrons sur les algorithmes supervisés, c'est-à-dire qui apprennent à partir de données labélisées (qui portent sur des personnes dont on sait si elles ont commis des infractions et, le cas échéant, lesquelles). Ce choix est motivé par le fait que ces algorithmes sont plus performants, et parmi les plus répandus (LeCun, 2016), et qu'il existe de la donnée pour les alimenter. Par ailleurs, nous nous intéressons

aux tâches de classification, qui donnent la probabilité qu'un élément appartienne à une certaine classe (par exemple la probabilité qu'une personne ait commis une infraction donnée – classe 1 – ou non – classe 0).

Prisme de perception : le sujet numérique effectif

Le sujet numérique prend deux formes singulières : une forme « diffuse », comme masse informe de données ; une forme « effective », opérationnalisée par un algorithme.

Le sujet numérique diffus ne constitue pas une information intelligible pour l'humain. Nous nous concentrons donc dans cette section sur le sujet numérique effectif, qui est le résultat d'un traitement algorithmique. Celui-ci se matérialise sous forme d'un score qui représente la probabilité qu'une personne commette telle infraction, étant donné la représentation numérique que l'on a de celle-ci et les données d'entraînement utilisées. L'algorithme se présente donc comme l'étape de distillation permettant d'obtenir l'essence de ce sujet diffus et dépend de deux types de données : celles récoltées sur la personne en particulier et celles qui ont servi à entraîner l'algorithme.

Il existe différents algorithmes de ML supervisés, et chacun de ces algorithmes transforme la donnée différemment, menant à des sujets effectifs différents à partir d'un sujet diffus unique. Nous proposons une première compréhension de ces différents sujets selon l'algorithme choisi, en nous concentrant sur trois d'entre eux particulièrement connus et dont les fonctionnements diffèrent fortement : la machine à vecteur de support (SVM), les arbres hiérarchiques et les réseaux de neurones artificiels (ANN pour *artificial neural network*).

- Le SVM est un algorithme de classification qui s'appuie sur deux grands principes. D'une part, il recherche une frontière telle que la distance significative des classes qui lui sont adjacentes soit maximale. D'autre part, comme les observations ne sont pas nécessairement séparables dans l'espace mathématique auxquelles elles appartiennent, cet algorithme permet de projeter celles-ci dans des espaces plus complexes dans lesquels la structure des données, c'est-à-dire l'existence de classes comme sous-espaces séparés, sera plus saillante. Ainsi, cet algorithme crée des sujets numériques effectifs dont les différentes opérations sur ceux-ci visent à les discriminer au maximum. Cependant, cette discrimination se fait dans un espace abstrait qui nous est inintelligible, au sein duquel certaines qualités du sujet sont alors « amplifiées ». Ces qualités se présentent cependant comme des abstractions mathématiques indépendantes de notre culture et inintelligibles à l'humain. La re-projection de ce résultat dans le monde physique pourrait mener à des discriminations difficilement explicables. Par exemple, sans préjuger de l'algorithme utilisé, l'utilisation d'un algorithme prédictif par un sheriff aux États-Unis a mené à des situations de harcèlement de personnes, sans raison apparente à nos yeux (Holmes, 2020).
- Les arbres hiérarchiques, simples ou complexifiés (forêts aléatoires, *boosting*) sont des algorithmes de classification, qui hiérarchisent les variables descriptives de la personne (donc ses données) et les découpent par régions de valeurs telles que les régions finales correspondent à des classes. Contrairement au SVM précédent, ces arbres attachent une grande importance aux variables constitutives du sujet numérique diffus. Celles-ci ne sont pas projetées, mais conservées et minutieusement découpées jusqu'à ce que l'on estime le résultat obtenu satisfaisant. Ces arbres supposent qu'il est possible de décrire par hiérarchisation de caractéristiques les individus, à travers leur sujet numérique effectif, et d'en tirer des conclusions quant à la nature plus ou moins criminelle de ces individus. Le logiciel COMPASS hiérarchisait aussi des dimensions (Angwin *et al.*, 2016), mais ici celles-ci sont créées à partir du sujet diffus, qui dicte donc les valeurs selon lesquelles un individu peut être classifié comme représentant ou non un danger, et un traitement algorithmique plus ou moins complexe. Nous proposons donc de nommer cette forme de sujet numérique « le sujet hiérarchisé ».

- Enfin, les derniers types d'algorithmes que nous abordons sont les réseaux de neurones artificiels (ANN). Ceux-ci ont connu ces dernières années un succès fulgurant, en particulier grâce à leurs performances impressionnantes dans les tâches de reconnaissance d'image. Ces algorithmes sont structurés par des « couches » : une couche d'entrée, à travers laquelle les données sont alimentées, une couche de sortie, qui correspond à la classification recherchée, et des couches intermédiaires. Chacune des couches intermédiaires a une fonction spécifique, attribuée automatiquement par l'algorithme, permettant de capturer par découpages successifs les différences topologiques existant entre les classes. Ces algorithmes, en particulier sous forme de réseaux profonds, ont tendance à mémoriser la forme des phénomènes plutôt qu'à chercher à en découvrir des structures explicatives (même dans un monde algorithmique), ce qui s'illustre par leur grand nombre de paramètres, par exemple le VGG19 est réputé pour ses très bonnes performances et contient 144 millions de paramètres. Ainsi, ils supposent qu'en appliquant un nombre de décompositions assez importantes, l'entité considérée sera presque « indifférenciable » d'une entité reconstituée (sous forme d'agrégat réticulaire de données de plusieurs entités), ce qui suppose une vision relativement déterministe et circonscrite du monde. Nous proposons donc de nommer ce troisième sujet effectif « le sujet circonscrit ». Par exemple, estimer qu'analyser une photo suffit à déterminer une personne, comme ce fut le cas pour Robert Williams, qui écopa d'une garde à vue de 30 heures (Le Monde avec l'AFP, 2020), ne représente-t-il pas une perception circonscrite ? Ce prisme de perception tend à ignorer la singularité du sujet, et soutient l'idée qu'il est toujours possible d'en déterminer une essence plus ou moins criminelle.

Sujet numérique diffus et effectif dans le droit pénal

Nous avons vu que les algorithmes de ML, alimentés par des données individuelles et comportementales, constituent des sujets numériques comme prismes d'observation des individus, pouvant servir d'indices à la construction future d'éléments de preuve, d'une part, ou pouvant servir à orienter l'enquête, d'autre part.

Dans un premier temps, ces sujets numériques, que nous nommons « diffus », sont constitués par une masse informe de données inintelligibles à l'humain, à cause de leur volume, structure ou mode d'agrégation. À ce stade, la matière d'analyse a été sélectionnée algorithmiquement et est aveugle à la question des dimensions conventionnelles ou culturelles d'analyse, qui devraient être, selon nous, privilégiées. Pour être clairs, ce sont les fournisseurs de données (réseaux sociaux, *data-brokers* ou autres) qui auront déterminé ce qui est à inclure ou non dans l'analyse, mais aussi les relations à prendre en compte dans ces données.

Dans un deuxième temps, lorsqu'un sujet « effectif » est construit, celui-ci se résume en la probabilité qu'il ait commis (selon l'ensemble des variantes possibles) un crime ou un délit. Ce sujet effectif est le résultat de diverses transformations qui impliquent que le sujet (réel) est observé à travers un prisme particulier, dont potentiellement personne n'est conscient. Nous avons caractérisé trois sujets effectifs spécifiques construits à partir d'algorithmes particuliers. Premièrement, le sujet « amplifié » par un algorithme de SVM, qui par abstraction mathématique dans des espaces complexes permet « d'amplifier » la discrimination entre divers sujets, amplification faite selon des règles inaccessibles à l'humain. Deuxièmement, le sujet « hiérarchisé » par des arbres hiérarchiques, où le sujet diffus segmenté et recombiné permet de déterminer les traits saillants d'un profil « criminel » ou non spécifique. Ces traits saillants, encore une fois, sont déterminés algorithmiquement, et rien ne permet de savoir si d'autres dimensions du sujet existent que l'algorithme n'estime pas importantes. Enfin, le sujet « circonscrit » par l'ANN, qui par zooms successifs tente d'en déterminer une essence déterminante. Cette essentialisation résorbe totalement le déterminisme de l'aléa (Longo, 2019).

Comprendre ces processus nous semble central, d'autant qu'ils pourraient modifier les frontières du droit pénal. En effet, jusqu'à présent, la responsabilité d'un individu face à la loi consiste en

la possibilité de déterminer avec assez de certitude si ce dernier a commis un acte délictueux ou criminel. Cela signifie, d'une part, que l'enquête porte sur une action et, d'autre part, que cette action se situe dans le passé. À l'inverse, l'utilisation du ML modifie le point de focalisation de l'action vers le sujet et du passé vers le futur, ou de la réalisation vers le potentiel criminel (puisque les données peuvent être ultérieures, ou être utilisées pour lier un comportement à un crime potentiel). Enfin, le sujet observé n'est pas le sujet physique, mais le sujet numérique. Ainsi, ces algorithmes incarnent une puissance de changement importante, qu'il convient de continuer d'explorer de manière à être conscients de leurs effets potentiels et à mettre en place les régulations nécessaires.

Bibliographie

ANGWIN J., LARSON J., MATTU S. & KIRCHNER L. (2016), "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*.

ARVIDSSON A. (2016), "Facebook and Finance: On the Social Logic of the Derivative", *Theory, Culture & Society*, 33(6), 3–23. https://doi.org/10.1177/0263276416658104

BELLO, I. & DAOUD E. (2020), "Les nouveaux moyens de lutte contre la fraude fiscale", *Revue Lamy Droit des Affaires*.

CHAMAYOU G. (2015), « Avant-propos sur les sociétés de ciblage », *Jef Klak*, 2, 1–12.

Code pénal — « Article 121-5 ».

CORNU G. (2020), *Vocabulaire juridique* (13e éd.) , PUF.

DEMIAUX V. & SI ABDALLAH Y. (2017), « Comment permettre à l'homme de garder la main ? », CNIL.

FOURCADE M. & HEALY K. (2013), "Classification situations: Life-chances in the neoliberal era", *Accounting, Organizations and Society*, 38(8), 559–572. https://doi.org/10.1016/j.aos.2013.11.002

GORIUNOVA O. (2019), "The Digital Subject: People as Data as Persons", *Theory, Culture & Society*, 36(6), 125–145. https://doi.org/10.1177/0263276419840409

GROSS J. (2018), "Speech By Lord Justice Gross Disclosure – Again", *Judiciary of England and Wales*.

HANSEN H. K., & FLYVERBOM M. (2015), "The politics of transparency and the calibration of knowledge in the digital age", *Organization*, 22(6), 872–889. https://doi.org/10.1177/1350508414522315

HOLMES A. (2020), "A sheriff launched an algorithm to predict who might commit a crime. Dozens of people said they were harassed by deputies for no reason", Business Insider. https://www.businessinsider.fr/us/predictive-policing-algorithm-monitors-harasses-families-report-2020-9

Le Monde avec l'AFP (2020), « États-Unis : Un Américain noir arrêté à tort à cause de la technologie de reconnaissance faciale », *Le Monde*.

LECUN Y. (2016), « Les Enjeux de la Recherche en Intelligence Artificielle. Leçon Inaugurale. Chaire Informatique et sciences numériques », Collège de France.

LONGO G. (2019), "Letter to Turing", *Theory, Culture & Society*, 36(6), 73–94. https://doi.org/10.1177/0263276418769733

WEXLER R. (2017), "The Odds of Justice: Code of Silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out", *Washington Monthly*. https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/