Algorithms and penal law: What does the future hold?

Elise Berlinski,
ESCP,
Imane Bello,
attorney,
&
Arthur Gaudron,
researcher, Centre de Robotique.

Abstract:
The algorithms of artificial intelligence (AI) and, in particular, machine learning (ML) are now penetrating the field of penal law. They create a numerical subject like a prism for perceiving the real subject to which they refer. The relation between the real subject and its digitized referent is not evident. Since the information provided by the latter can orient an investigation or influence a judge's opinion, it is urgent to much better understand the filters through which these methods let us perceive the physical person under investigation or standing before a judge. The stages of the construction of the digitized subject, ranging from the data used to algorithmic applications, are described. These tools and techniques create kaleidoscopic reflections. What changes does the creation of the digitized subject via ML algorithms imply for penal law?

> "*Proprietary algorithms are flooding the criminal justice system. Machine learning systems deploy police officers to 'hot spot' neighborhoods. Crime labs use probabilistic software programs to analyze forensic evidence. And judges rely on automated 'risk assessment instruments' to decide who should make bail, or even what sentence to impose*" (WEXLER 2017).

The algorithms used for artificial intelligence (AI), specifically machine learning, are now a reality in the American penal system and will soon come to France, evidence of this figuring in the CNIL's report, which discusses their application in penal law (DEMIAUX & SI 2017). Herein, I would like to focus on the implications and changes ensuing from the use of machine learning in the field of French penal law.[1]

The application of machine learning algorithms in penal law is intended to produce "clues" for the purpose of helping to determine how a reprehensible act actually happened, or to orient an investigation. The use of such algorithms forces us to ask questions about both the data they are fed and the calculations they regurgitate. These data, which increasingly come from the social media (GROSS 2018, p.8), are often behavioral, and have been de- and re-aggregated. They are used to "construct a subject" that is normally treated as being representative of a real person. The data contain this subject's "essence" (CHAMAYOU 2015). As studies have shown however, there is no evident match between this representation of a "digital subject" and the real person serving as referent (GORIUNOVA 2019). It is, therefore, urgent to understand the filters of these "*mediating technologies*" (HANSEN & FLYVERBOM 2015) and their effects when technology serves to make a representation of a person.

After a brief presentation of the penal system in France, we shall describe the methods used to construct a "digital subject" and then draw a few conclusions.

[1] This article, including any quotations from French sources, has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references. All websites were consulted in April 2021.

Evidence and legal arguments under penal law

In penal law (CORNU 2020), an offense is qualified by bringing together the following elements: the specification of the articles of law relative to the offence (the legal element), a behavior that is related by causality with an effect (the material element) and the free consent for taking part in an act that was known to be illegal (the moral element). Offenses may be material (with damages) or immaterial. In the latter case, the accomplishment of the incriminated act by itself suffices for qualifying it as an offense (*e.g.*, counterfeiting money). Furthermore, an attempt to commit a crime is reprehensible if "*it has been suspended or has been lacking in effect only owing to circumstances independent of the perpetrator's intent*" (Article 121-5 of the Penal Code).

Police officers could, we imagine, use machine learning algorithms to determine the probability of the occurrence of a crime or offense, and then launch an investigation to know whether it actually occurred. For instance, the fiscal administration[2] uses such algorithms for detecting inconsistencies that might be qualified as fraud or be low-level signals of tax evasion (BELLO & DAOUD 2020).

For an eventual conviction, a judge has to analyze the qualification of an offense that happened in the past and was embodied in a behavior. This analysis is based on evidence, which, under penal law, is "free". Under Article 427, "*apart from the cases where the law has foreseen otherwise, offenses may be proven by any kind of evidence, and the judge makes a decision based on his deep-seated convictions. The judge may base his decision only on the evidence presented to him during hearings and discussed and cross-examined in his presence.*" So, the results of using statistical models or search systems using algorithmic data, may not, strictly speaking, be used as evidence of the realization of an act but might, nevertheless, influence the judge's sovereign appreciation of a case.

Although there are no legal grounds for taking action against an act that has not been accomplished (materially, immaterially or as an attempt), the output from algorithms might influence judges (and their perception of reality) and/or orient investigations. Even though the algorithmic results cannot be used without tangible proof of the execution of the act, we can imagine that they might weigh on sentencing and on the formation of the judge's personal convictions. According to the principle of law on the individualization of sentencing, the personality and singularity of the accused are to be taken into account when a sentence is pronounced. The second effect of algorithms might be to orient the decision about conducting an investigation (and about how to conduct it). In all cases, it is necessary to understand the filters used by this technology in order to gauge this phenomenon.

Constructing a digital subject

Two blocks are used to build a "digital subject": data and the algorithmic calculations for determining the probability of whether or not an individual committed an offense. The data describing the subject might come from the legal system (all the records related to a case, ranging from declarations to the police to court decisions) and other sources, public or private (information collected by electronic sensors or on the social networks). This profile is compared (by a model) by using databases of observations or sets of training data with information about the offenses committed by a person and the additional data digitally collected from various sources. At present, the social networks are "good" databases, since they have cells all over the Web for forming behavioral profiles that are then sold to the organizations that want to assess risk profiles (for loans, insurance, etc.) (ARVIDSSON 2016, FOURCADE & HEALY 2013).

[2] Both the Direction Générale des Finances Publiques and the Direction Générale des Douanes et Droits Indirects.

The algorithms used for calculations do not stem from a theory (*e.g.*, sociological). Instead, they are trained through a learning process based on data and statistical inferences. We shall concentrate on the supervised algorithms that learn from labeled data; these data concern persons who are known to have (or not have) committed offenses and, if so, which ones. The reason for this choice is that these algorithms are already operational; they figure among those the most widely used (LECUN 2016). Furthermore, there are data to feed them. We are interested in the tasks of classification that assign probabilities to elements belonging to a given class (*e.g.*, the probability that a person has committed a given offense or not).

Filtered perceptions: The diffuse and effectual digital subjects

The digital subject has two distinct forms: the one "diffuse" like an unformed mass of data; the other "effectual", *i.e.*, made operational by algorithms. Since the diffuse subject does not correspond to information intelligible to human beings, we shall concentrate on the effectual subject, which is the output of algorithms. This subject is materialized as a score corresponding to the probability that a person will commit an offense given his/her digital profile and the training data that have been used. Algorithms are, therefore, used to "distill" the diffuse digital subject's "essence". They depend on two types of data: those harvested about the person in question and those used to train the algorithms.

Supervised machine learning uses various algorithms. Each algorithm operates on data and produces effectual subjects out of the single diffuse subject. In a first approach to understanding these subjects in relation to the algorithms, we propose concentrating on three particularly well known algorithms with quite different functions: support-vector machines (SVM), hierarchical tree structures and artificial neural network (ANNs).

The SVM is an algorithm of classification based on two major principles. For one thing, it seeks to establish a boundary such that the gap between the classes adjacent to the boundary is maximal. For another, since observations cannot necessarily be separated into classes in the mathematical space where they belong, this algorithm projects them into more complex spaces in order to structure the data and make the existence of classes as separate subspaces more salient. This is how SVM creates effectual digital subjects. Its various operations on these subjects seek to discriminate maximally between them. However this discrimination is made in an abstract space unintelligible to us, where certain characteristics of the subject are "amplified". These characteristics are presented as mathematical abstractions independent of culture and unintelligible to human beings. The re-projection of these results in the material world might, therefore, lead to discriminations that are hard to explain. For example (and without judging SVM as such), the use of a predictive algorithm by a sheriff in the United States led to the "*serial harassment*" of people for no apparent reason (HOLMES 2020).

Hierarchical tree structures, whether simple or complex ("boosted"), are algorithms of classification that rank the variables (and thus the data) used to describe a person and divide them into sets of values such that the final sets correspond to classes. Unlike in SVM, these trees place importance on the variables that constitute the diffuse digital subject. Instead of being projected, these variables are stored and analytically broken down until a satisfactory result is obtained. Underlying these trees is the supposition that it is possible to hierarchically rank an individual's characteristics by using his/her effectual digital subject and to draw conclusions about the propensity to commit an offense. COMPAS software has been used to make these dimensions (ANGWIN *et al*. 2016), but now the dimensions are created out of the diffuse digital subject, which dictates the values whereby an individual can be classified as representing a danger or not, following an algorithmic process of varying complexity. We have called this sort of digital subject the "hierarchized subject".

As for artificial neural network (ANNs), their success in recent years has been stupendous thanks to their impressive performance in image recognition. These algorithms are structured in "layers": an input layer where the data are fed, an output layer with the sought-for classification, and, between the two, several other layers. Each intermediate layer has a specific function that, automatically assigned by the algorithm, performs successive operations to detect topological differences between classes. These algorithms, in particular deep neural networks, tend to memorize the form of phenomena rather than trying to discover structures that would be explanatory (even in the language of algorithms). Evidence of this is the large number of parameters (*e.g*, VGG19, which is known for its performance, has 144 million parameters). The supposition is that, by enough successive applications of the algorithm, the entity under study will nearly be "undifferentiatable" from the reconstructed entity (in the form of an aggregate of data from several entities). Since this implies a rather deterministic vision of a world that can be circumscribed, we refer to this third effectual data subject as the "circumscribed subject". Is perception not circumscribed when, for example, we assume that analyzing a photograph suffices to determine who someone is — as happened to Robert Williams, who was erroneously held in custody for thirty hours (LE MONDE & AFP 2020)? This filter on perception tends to fail to detect the subject's singularity and sustains the idea that it is always possible to determine an "essence" that is more or less "criminal prone".

These digital subjects in penal law

As we have seen, machine learning algorithms that are fed data about individuals and behaviors construct digital subjects as prisms that filter our perceptions of individuals. The "evidence" they provide could be used either to prove a case or orient an investigation.

Initially, the "diffuse" digital subject is an amorphous mass of data unintelligible to human beings because of their volume, structure or aggregation. During this first phase, algorithms select what is to be analyzed. They are blind to questions about the usual or cultural dimensions of analysis that, in our opinion, should be given priority. To be clear, it is the data-providers (social networks, data-brokers, etc.) who will determine not only what is to be included or not in an analysis but also the relations in the data to be taken into account.

In a second phase, once an "effectual" digital subject has been constructed, it is represented by the probability that this subject has committed (depending on the set of possible variants) a crime or misdemeanor. This effectual subject is the outcome of various transformations that imply that the "real" subject is being viewed through a filter. Potentially, no one is aware of this filtering process.

We have described three sorts of effectual subjects constructed by algorithms:

- the "amplified" subject that is the output of SVM, which, through mathematical abstraction in complex spaces, amplifies the discrimination between subjects by using rules that are inaccessible to human beings.
- the subject as "hierarchized" by using tree structures. The diffuse digitial subject is segmented and recombined so as to determine whether the salient traits of a criminal profile apply or not. In this case too, algorithms define the traits with no way for us to know whether other dimensions of the subject exist that the algorithms have not deemed important.
- the subject as "circumscribed" by ANN, which, through successive zooms, tries to determine the subject's essence, a process that fully reabsorbs the determinism of contingencies (LONGO 2019).

We think that it is crucial to understand these processes, even more so when they risk shifting bounds within penal law. Till present, an individual's responsibility in the eyes of the law has lain in the possibility of accurately enough determining whether he/she has committed an offense. This implies both that an investigation focuses on an action and that this action lies in the past. In contrast, the use of machine learning shifts the focal point from the action toward the subject and from the past toward the future, or from the realization of an offense toward the potentiality of being a criminal. After all, the data might come from a date after the act or be used to link a behavior to the potential for crime. Finally, the subject thus "observed" is not the physical subject; it is the data subject. These algorithms thus have the power to wreak major changes, which we should continue exploring in order to be aware of their potential impact and adopt the necessary regulations.

References

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*, 23 May, available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

ARVIDSSON A. (2016) "Facebook and finance: On the social logic of the derivative", *Theory, Culture & Society*, 33(6), pp. 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* (Alphen-aan-den-Rijn, NL: Wolters Kluwer).

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

CORNU G. (2020) *Vocabulaire juridique*, 13th edition (Paris: Presses Universitaires de France).

DEMIAUX V., & SI A.Y. (2017) *Comment permettre à l'homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, December (Paris: Commission Nationale de l'Informatique et des Libertés), 80p., available via https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf.

FOURCADE M. & HEALY K. (2013) "Classification situations: Life-chances in the neoliberal era", *Accounting, Organizations and Society*, 38(8), pp. 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), pp. 125-145, available at https://doi.org/10.1177/0263276419840409.

GROSS J. (2018) "Speech By Lord Justice Gross: Disclosure – Again" (Judiciary of England and Wales), 9p., available via https://www.judiciary.uk/wp-content/uploads/2018/06/lj-gross-disclosure-speech-june18-1-1.pdf.

HANSEN H.K., & FLYVERBOM M. (2015) "The politics of transparency and the calibration of knowledge in the digital age", *Organization*, 22(6), pp. 872-889, available via 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*, 10 September, available at https://www.businessinsider.fr/us/predictive-policing-algorithm-monitors-harasses-families-report-2020-9.

LE MONDE & AFP (2020) "États-Unis: Un Américain noir arrêté à tort à cause de la technologie de reconnaissance faciale", *Le Monde*, 24 June, available. At https://www.lemonde.fr/international/article/2020/06/24/un-americain-noir-arrete-a-tort-a-cause-de-la-technologie-de-reconnaissance-faciale_6044073_3210.html.

LECUN Y. (2016) "Recherches sur l'intelligence artificielle. Qu'est-ce que l'intelligence artificielle?" 7p., available at https://www.college-de-france.fr/site/yann-lecun/Recherches-sur-l-intelligence-artificielle.htm.

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

WEXLER R. (2017) "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*, June/July/August, available at https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/.