THE ROLE OF PREDICTIVE SYSTEMS IN CRIMINAL INVESTIGATIONS: ACCOUNTABILITY, DISCRIMINATORY EFFECTS AND FAIR TRIAL GUARANTEES

Giuseppe Contissa & Giulia Lasagni
Alma Mater Studiorum - University of Bologna

Today ICT systems are supporting the progressive automation of almost all the human activities and processes related to information. According to Balkin, we are already living in an algorithmic society, that is a society ``organized around social and economic decision making by algorithms robots, and artificial intelligence agents; who not only make the decisions but, in some cases, also carry them out'' (BALKIN, 2016).
In the last few years, we have seen an explosion of systems developed on the basis of the (assumed) predictive capabilities of algorithms applied on Big Data: «Originally created to understand consumer behavior – will a person who buys product X also buy product Y? – big data analytics has increasingly come to be seen as the solution to any problem involving large amounts of data» (MILLER, 2014).
These systems adopt data mining technique for discovering useful patterns (set of relationships) in large data sets, using machine learning algorithms and statistics, that are subsequently employed for algorithmic decision-making: On the basis of the discovered relationships, the systems learn how to classify individuals into target classes, or how to assess the values of a target variables (*e.g.*, low/high/medium individual recidivism rate; low/high/medium crime risk in a particular geographic area). Such target values may express prediction concerning the behaviour of the concerned individuals, or group of individuals, or areas, and elaborates strategies on how to treat with those risks on the basis of their classification.

In criminal justice, algorithmic systems are reported to be used at least for the following purposes: Predicting crimes; predicting offenders; predicting perpetrators' identity; and predicting victims (cf. PERRY ET AL., 2013). In doing so, two different approaches may be used: A first one, mirrors conventional crime mapping and investigative methods, basically supporting or substituting human investigative experience with an integrated analysis of already available data to discover potential criminal patterns. These systems for predictive justice, which appear rather helpful for a better allocation of resources in criminal justice, should be able to predict the probability that a certain crime will be committed, where, and when, on the basis of the analysis of data concerning social, demographics, economic, environmental factors, and data concerning previously committed crimes (that is, for instance, the case of PredPol, a predictive algorithmic system developed by the Los Angeles Police Department and behavioral scientists of UCLA; or of KeyCrime, developed by the Milan Police Department).
A second approach, that appears much more critical in light of fundamental rights protection, uses predictive analytics methods that, accessing to huge amount of data (not necessarily already available to law-enforcement), correlate risk factors with specific individuals thanks to mathematical models that autonomously identify individuals as, for instance, potential offenders (that is, for instance, the case of COMPAS, developed

by a Californian private company, and used to predict individual recidivism risk, cf. SKEEM AND LOUDEN, 2007).

The use of such algorithmic systems for criminal justice purposes, and in particular for criminal investigations, represents an incredibly challenging element, potentially able to standardize the criminological evaluation on single individuals, but also to generate great distortions in the fairness of criminal proceedings as a whole, especially in those (most) legal systems where the impact of new technology, and predictive systems in particular, as not yet been into due account at the legislative level.

A first detrimental effect of the predictive algorithmic systems, is represented by discriminatory effects.

The analysis and decisions taken by computers in fact often enjoy an undeserved assumption of fairness or objectivity (KROLL ET AL., 2016). This approach has been defined *data fundamentalism* (CRAWFORD, 2013), namely the tendency to believe that the correlation assessed by the algorithm implies causality, and that the analysis carried out with data mining techniques on large sets of data always provides an objective view of reality. This is a particularly dangerous assumption, especially in cases in which the analysis is aimed at obtaining predictive information of behaviours, intentions, and attitudes of individuals. Instead, the design and implementation of these automated decision systems can be vulnerable to a variety of problems that can result in systematically faulty and biased determinations (KROLL ET AL., 2016).

Firstly, algorithms that include some type of machine learning can lead to discriminatory results if they are trained on historical examples that reflect past prejudice or implicit bias. For example, let's assume that a program to identify suspects is trained on the previous human law-enforcement decisions, and those previous decisions were themselves racially biased, so that Middle-Easterners or Afro-American people were more likely to be categorized as suspects. This would result in the system to reproduce biased or discriminatory results, on the basis of algorithmic procedures that are apparently objective, but are affected by the bias "inherited" from previous human decisions.

Secondly, data may offer a statistically distorted picture of groups comprising the overall population, leading again to discriminatory results. For example, let's assume that an algorithm that instructs police to stop and control pedestrians is trained on a dataset that over-represents the incidence of crime again among certain ethnic groups. This would cause the algorithm to instruct police to stop and control more people belonging to these groups than from others, with the results that statistically, more new crimes committed by the selected ethnical group will be discovered than those committed by other groups. Then, when data concerning the new crimes will be added to the dataset, it will even more over-represents crimes among those groups, reinforcing the discriminatory effect, in a Catch-22 situation (MILLER, 2014).

The use of this practice in the EU would appear also to be in breach of Article 11 Directive 2016/680, according to which decision based solely on automated processing, including profiling that results in discrimination against natural persons on the basis of special categories of personal data (revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic

data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation) shall be prohibited.

Another legal challenge, very critical in terms of transparency and accountability, is represented by data protection, and it is linked to the activities of data collection and profiling.
Algorithmic decision-making requires first of all to collect data about individuals. Then, automated classification (profiling) extracts new data about the concerned individuals. The use of such systems may even lead to discover sensitive data from non-sensitive ones (Famous the case of the AI data mining system deployed at Target (a supermarket and discount store chain), that in 2012, on the basis of the analysis of purchasing data (including unscented lotion, mineral supplements, and cotton balls) correctly inferred that one of its customers, a teenage girl, was pregnant).
Although it might be argued that citizens themselves have given up much of their personal information to third parties, the possibility for private companies, in addition to public authorities to record, store, and analyse nearly everything they do it is not to underestimate, as it implies business companies and law-enforcement actors gaining more power, and individuals losing a measure of liberty.
In this sense, the possibility that potentially undisclosed conflicts of interests (*e.g*. of commercial nature) might have a role in determining the selection process applied by the algorithm should also be taken into account, given that most of predictive software is developed and owned by private companies for profit purposes (that is, for instance, the case of COMPAS).

Critical issues in the use of algorithmic systems for criminal justice purposes also emerge with regard to profiles specifically linked to criminal investigations, first of all the presumption of innocence, which plays an essential role in ensuring that proceedings against individuals are effectively in compliance with the principle of human dignity, and with the other procedural fundamental rights (see, *e.g.*, QUINTARD-MORÉNAS, 2010; TRESCHEL, 2005; ILLUMINATI, 1979). The presumption, enshrined in most international legal sources dealing with human rights protection (and in particular by Article 6(1) and (3) ECHR, and Article 48 of the Charter of Fundamental Right of the European Union-CFREU), although often differently protected at the national level, lies at the heart of the notion of fair procedure developed by the Court in Strasbourg and adopted by the Court of Justice (see, among the others, ECtHR, *Barberà, Messegué and Jabardo v. Spain,* 6 December 1988, Application no. 10590/83; *John Murray v. the UK*, 28 October 1994, Application no. 18731/91; ECJ, *Telefónica e Telefónica de España v. Commission*, Case T-336/07, 29 March 2012, ECLI:EU:T:2012:172; *Criminal proceedings against Marcello Costa and Ugo Cifone*, Joined Cases C-72/10 and C-77/10, 16 February 2012, ECLI:EU:C:2012:80).
In this perspective, the use of predictive systems seriously threatens the right of the defendant not to be presented as guilty before proven so according to law, which is a right, it shall be reminded, that does apply also during the pre-trial phase. Indeed, while the presumption does not prevent judicial authorities to apply provisionary measures based on prognostic judgement, the latter cannot get to the point of anticipating a verdict

of culpability, which is what happens when automated predictive systems are used to support decision-making process in time constraints contexts (*e.g.* Frontex border control decisions in migration flows, cf. Fergusson and o., 2014), or create presumptions which are, by law or fact, irrebuttable.
Also this practice, potentially applicable in the EU, would appear to be in breach of Article 11 Directive 2016/680, according to which decision based solely on automated processing, which produces an adverse legal effect concerning or significantly affecting a subject, shall be prohibited unless authorized in (clear) legal basis which provides appropriate safeguards for the rights and freedoms of the latter, especially if based on the special categories of personal data mentioned above.

A further, and potentially even more serious, critical issue in the use of predictive algorithmic systems for criminal justice purposes concerns the very essence of criminal investigation itself, that is being a procedure characterized by the use of invasive investigative powers on the basis of democratically established legal criteria, and under the supervision (during, or ex post) of a judicial authority.
Indeed, predictive systems, similarly to digitally advanced form of surveillance (such as the use of drones, or malware in communication or other kind of data interception), are contributing to dramatically blurring the line between preventive investigations and criminal (that is, post-factum) inquiries. In theory, the difference between the two is rather straightforward: The first answer to fundamental public interests, such as national security, are generally poorly regulated (or regulated with a very flexible legal basis), supervised only by political or administrative authorities, and the elements so gathered may be used only for intelligence purposes. On the other side, criminal investigations provide for a careful balance of powers and guarantees, enshrined in the law (or in foreseeable case-law), which shall be respected in order for the evidence so obtained to be used at trial.
This distinction, however, becomes less clear when relevant amounts of data are collected not during criminal investigations, but by administrative authorities, or even more significantly, by private operators acting for administrative purposes, as it is the case of predictive algorithmic systems. The use of these systems, in a globalized context in which information are likely to circulate in horizontal, vertical and diagonal forms of cooperation, and where no harmonized exclusionary rule has been established at transnational level, results especially worrisome. In these cases, in fact, elements and information, relevant also for criminal proceedings, may get collected in spite of the procedural guarantees provided for in criminal investigations, potentially including judicial authorization. This critical issue currently remains unsolved in the EU, either at the legislative level, and in the ECJ case-law (although some criteria may de deduced by the 2015 decision *WebMindLicence* which dealt with the opposite case of the use of evidence gathered in criminal investigations within administrative proceedings; see TESORIERO, 2016).
The mentioned critical effects appear all the more severe taking into account that the lack of clear legal basis (if any) in which predictive algorithmic systems operates, combined with the limited access to the concrete functioning of the algorithmic system (if any), *de facto* generates a situation in which the person addressed by adverse measures due to the use of such predictive systems often is not able to exercise any effective remedy

for complaining about violations of her/his rights and freedoms, a situation which clearly appears as an infringement of both Articles 13 ECHR and 47 CFREU. This lack of remedies needs to be solved also in light of Directive 2016/680, which at Recital (104) requires that also the processing of personal data for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, shall be in compliance with the right to an effective remedy and to a fair trial, and that limitations to the latter may be accepted only as long as they are in accordance with Article 52(1) CFREU.

Against these critical issues, clearly optimal solutions are not straightforward (given that the exclusion of digital technology from the criminal justice systems does not seems a feasible option, being both anachronistic and impractical (not to say impossible) to realize).

With specific regard to data protection in the EU, there a pressing need emerges to find a way to govern the information system, monitor its automated data processing activities, review its decisions and correct those that are incorrect or unfair.

The first goal when designing an automated decision system therefore should be how to ensure transparency and accountability: It should be possible to demonstrate that rules constituting a policy were correctly implemented in the system, and applied evenly across all subjects, but also that those subjects can be certain that the policy is designed to promote substantive goals or principles of the law. According to Recital (71) GDPR, «in any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision». However, to date it is not clear what does explanation or transparency mean, and what form they should take.

Perhaps the most obvious approach is to disclose a system's source code, but delivering the source code may be limited first of all by intellectual property issues. Even when the source code is disclosed, this is at best a partial solution to the problem of accountability for automated decisions. In fact, the source code of computer systems is usually illegible to non-experts, and even experts often struggle to understand what software code will do. Therefore, inspecting source code is a very limited way of predicting how a computer program will behave. Moreover, machine learning involves situations where the decisional rule itself emerges automatically from the specific data under analysis, sometimes in ways that no human can explain. In this case, source code alone explains very little, since the code only exposes the machine learning method used and not the data-driven decision rule: The model is effectively a "black box" both for the concerned individuals and the experts.

A second approach, rather than providing the source code, would favour the disclosure of higher level information about the logic involved in the automated decision making. This is the approach followed also by the GDPR in Article 13.2(f), that identifies the information to be provided where personal data are collected from the data subject (the same provisions are repeated in Article 14.2(g) in relation to data not obtained from the data subject).

According to this approach, information to be disclosed may include the following: (1) information about the data that served as the input for automated decision; (2) information about the list of factors that influenced the decision; (3) information on the relative importance of factors that influenced the decision; and (4) a reasonable explanation (possibly in textual form) about why a certain decision was taken. But even the explanation of the functioning of the system may also be subject to some issues, as it is not clear whether they should include the disclosure also of personal data (concerning all the affected individuals) which have been used to reach the decision, the resulting determinations, or also the more complex explanation of how much each input data impacted on the final classification o determination.
In this context, a first issue may be reconnected to those cases in which the explanation would require the disclosures of the assessment (and therefore of personal data) of third parties, to compare the decisions taken; a possibility that is subject to limitations by the GDPR itself. Besides, the disclosure of such information may allow the system to be cheated: For example, in a system for the automated detection of tax evasion, the disclosure of the rationale adopted by the system to associate the reaching of specific thresholds values in tax declarations, to the detection of tax evasion, may favour the adoption of "strategic behaviours" by individuals when presenting tax declarations.

Rather than transparency, several experts suggest the adoption of *procedural regularity* (Kroll, 2016), namely the idea that decisions are reached in accordance with consistent and agreed upon rules, so that each individual knows that the decision was not made to discriminate her on purpose. To this end, engineers may adopt specific techniques so that the system can demonstrate compliance with key standards of legal fairness for automated decisions, without revealing key attributes of the decisions or the processes by which the decisions were reached.

With regard to the consequences more strictly brought by predictive algorithmic systems to criminal investigations, as highlighted by the Court of Justice in Joined cases *Tele 2 Sverige-Watson*, national legislation governing the protection and security of traffic and location data shall be precluded from granting access to retained data where the objective pursued by that access, in the context of fighting crime, is not restricted solely to fighting serious crime, where access is not subject to prior review by a court or an independent administrative authority, and where there is no requirement that the data concerned should be retained within the European Union. However, when such criteria need to be put into practice, the difficulty of retaining data before the beginning of specific criminal investigations, without creating important discriminatory effects (for instance, retaining data of all individuals belonging to a certain ethnicity or religion for terrorism purposes) clearly emerges.
On the other side also the alternative scenario, that is a generalized collection of data of all individuals, or all means of electronic communication or traffic data without any differentiation, possibly seems equally unpleasant, besides for having already been declared disproportionate and in violation of the rights of the EU Charter by the Court of Justice in the famous landmark *Digital Rights Ireland*.

Against this critical background, the paper concludes discussing some proposals and guidelines which may be applied to ensure a (at least better) fair balance between security and safeguards, in a democratic society, trying to apply the parameters developed by the two European Courts under Articles 7 and 8 CFREU and Article 8 ECHR to the matter of predictive algorithmic systems.

SHORT BIBLIGRAPHY

BALKIN, *The three laws of robotics in the age of big data*, in *Ohio State Law Journal*, 78, 2017

CRAWFORD, *The hidden biases in big data*, in *Harvard Business Review*, 1, 2013

EUROPEAN AGENCY FOR THE MANAGEMENT OF OPERATIONAL COOPERATION AT THE EXTERNAL BORDERS OF THE MEMBER STATES OF THE EUROPEAN UNION, and FERGUSSON, *Twelve Seconds to Decide: In Search of Excellence: Frontex and the Principle of Best Practice*, Publications Office of the European Union, 2014

ILLUMINATI, *La presunzione d'innocenza dell'imputato*, Zanichelli, 1979

KROLL, BAROCAS, FELTEN, REIDENBERG, ROBINSON, YU, *Accountable algorithms*, U. Pa. L. Rev., 165, 633, 2016

MILLER, *Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm*., in *J. Tech. L. & Pol'y*, 19, 105, 2014

PERRY, MCINNIS, PRICE, SMITH, HOLLYWOOD, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*, RAND Safety and Justice Program, 2013

QUINTARD-MORÉNAS, *The Presumption of Innocence in the French and Anglo-American Legal Traditions*, in *The American Journal of Comparative Law*, LVIII, 2010, 1, p. 107

SKEEM, LOUDEN, *Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*, 2007, available online

TESORIERO, *Processo penale e prova multidisciplinare europea in materia di illeciti finanziari*, in *Riv. dir. proc*., Anno LXXI (Seconda Serie) - N. 6, p. 1540, Novembre-Dicembre 2016

TRESCHEL, *Human Rights in Criminal Proceedings*, Oxford University Press, 2005