When it is (also) algorithms and AI that decide on criminal matters: In search of an effective remedy

WHEN IT IS (ALSO) ALGORITHMS AND AI THAT DECIDE ON CRIMINAL MATTERS: IN SEARCH OF AN EFFECTIVE REMEDY

Giulia Lasagni, Giuseppe Contissa*

European Journal of Crime, Criminal Law and Criminal Justice (forthcoming)

Abstract
The paper presents the main areas of application of predictive systems based on algorithmic and AI technology, and analyses their impact on fundamental rights and fair trial principles. It focuses in particular on the definition of the right to an effective remedy against decisions taken (also) with the support of algorithmic and AI systems, and proposes some innovative solutions on how to ensure compliance with this right in technologically advanced criminal proceedings.

Key Words: AI; algorithms; effective remedy; fair trial; transparency; discrimination; black box

1. Identifying Basic Notions and Concepts

Algorithms and artificial intelligence are progressively transforming almost all human activities, and in particular decision-making processes, making them dependent on their ability to record and process information. According to Balkin we are already living in an algorithmic society, that is, a society organized around the automated decision-making process, in which algorithms and artificial intelligence (A/AI) make decisions.[1]

On top of this progressive social automation described by Balkin there has also come, in recent years, a paradigm shift in AI that has led to the adoption of new methods based on knowledge induced by data analysis. Traditional computer-based decision support systems, which use human specialist knowledge, transferred into the system through symbolic representations of knowledge and logical inference, have

---

* The article is a joint reflection of the authors, but its drafting is broken down as follows: Contissa §§ 1, 2, 4.1, 4.2, 6 (first part, technical-informatics proposals); Lasagni §§ 3, 4, 4.3, 5, 6 (second part, legal proposals).

[1] J. M. Balkin, "The Three Laws of Robotics in the Age of Big Data," 78, *Ohio State Law Journal* (2017), 1219.

been integrated or replaced by AI systems based on *machine learning*, applied to large masses of data (so-called big data).[2]

Rather than carrying out evaluations and assessments on the basis of an algorithm containing a set of rules predefined by the programmer, the system builds its own model of the domain, applying a learning algorithm to analysis of the training data. Using this model, the system generates classifications, evaluations, and predictions on new cases submitted to it. Updating and expanding the dataset automatically improves the model and the system's predictive capabilities.

An A/AI system based on machine learning may provide better performances than systems based on symbolic approaches, but neither its functioning nor the reasons behind each decision can be fully explained by means of the source code, as that would only explain the operation of the learning algorithm, but not the final configuration of the model created by the system itself, which is the basis of its operations and decisions.

In this sense, these systems can be considered as *black boxes*, i.e., systems in which input and output are observable, while their internal functioning remains obscure even to their own programmers.[3] Their functioning therefore resembles that of an "oracle," but, contrary to its ancient predecessors, of an oracle with great statistical precision.

Criminal law is not immune to these transformations: A/AI systems are increasingly being introduced at different stages of the proceedings, supporting investigations (*predictive policing*) or court decisions (*predictive justice*).[4] As will be described in the following sections, these technologies can contribute to improving the efficiency of justice; but at the same time, their use also raises a number of concerns about the protection of fundamental rights.[5] In this article, we specifically focus on one of such profiles that bears major implications in criminal proceedings, trying to answer to the following research question: What can be considered an effective remedy against (totally or partially) automated decisions?

---

[2] "Machine learning" systems improve their performance by automatically learning how to perform future tasks by observation, see S. J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Englewood Cliffs, NJ: Prentice Hall, 2010).

[3] J. Millar, I. Kerr, "Delegation, Relinquishment, and Responsibility: The Prospect of expert robots," in *Robot Law*, R. Calo, A. M. Froomkin, I. Kerr, eds., Edward Elgar Publishing, 2016, 102–28, 107; F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2015).

[4] "Can you foresee a day […] when […] artificial intelligences, will assist with courtroom fact-finding or, more controversially even, judicial decision-making?" "It's a day that's here" (chief justice John Roberts, US Supreme Court) A. Liptak, "Sent to Prison by a Software Program's Secret Algorithms," *New York Times*, 1.05.2017.

[5] See, e.g., A. Garapon, J. Lassègue, Justice digitale: révolution graphique et rupture anthropologique (Paris: PUF, 2018).

In order to correctly frame the problem, we first present the most renown applications of A/AI systems in criminal justice, as well as the main criticalities they have raised. Although the focus of this paper is mainly on the decision-making phase, relevant issues concerning investigation and evidence will also be mentioned as far as they are necessary to understand the magnitude of the actual or potential changes deriving from the use of these technologies in criminal proceedings. Then, we assess the specific problems regarding the right to an effective remedy, when decisions are taken (totally or in part) with the support of such A/AI systems. To this goal, we compare some critical aspects of such automated decision-making with totally human one. We conclude with some legal and technological proposals on how a remedy should be structured in this context to be really considered effective.

Given the current lack of specific case law and legislation in the EU context, in the paper reference will be largely made to the US experience. This analysis, however, is very relevant also on our side of the Atlantic, where – against a relatively strong protection of fundamental fair trial rights, including that to an effective remedy – the debate on whether and how to rely on A/AI systems is still wide open.[6]

2. A/AI systems in criminal justice

Different criminal justice systems around the world today make use of A/AI systems to support the human decision-making of different actors, such as law enforcement agencies, lawyers, and judges.[7]

The basic scenario may be described as follows: the system classifies individuals into reference classes. These classes may express predictions about the behaviour of individuals, or groups of individuals (e.g., low/high/medium individual recidivism rate; low/high/medium risk of crime in a particular geographical area). These predictions are then used in the algorithmic decision-making process, i.e., to elaborate and suggest strategies on how to treat such individuals according to their classification.[8]

In this contribution we will focus exclusively on the use of A/AI systems in order to identify potential crimes or offenders; or—on a much more critical hypothesis, especially with a view to an effective remedy—in order to formulate individualised risk predictions.

---

[6] The considerations carried out in this paper appear relevant both for common law and civil law jurisdiction, however a detailed analysis of the potential specific differences of these systems goes beyond the remit of this contribution.

[7] W. L. Perry, B. McInnis, C. C. Price, S. Smith, J. S. Hollywood, *Predictive Policing: The Role of Crime Prediction in Police Operations* (Santa Monica, CA: RAND Corporation, 2013).

[8] Critical C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (New York: Crown Publishing Group, 2016), 10.

2.1 Identification of potential crimes and offenders

A first way to use A/AI systems for preventive purposes is to apply them to support (or replace) human investigative experience with an integrated analysis of available data, in order to identify potential crime patterns and reduce victimisation in digital environments, such as social media.[9] Another approach is to predict the circumstances (time and place) of possible offences. This approach reflects traditional investigative methodologies for mapping criminal activities in a given area by analysing data on social, demographic, economic, environmental, and criminal background.

There are several examples of these "more conventional" predictive police instruments used in Europe and the United States. Among the best known is PredPol, an algorithmic system developed by the police and the University of California, Los Angeles (UCLA). This system, based on historical data on offences (in particular those relating to victims), makes forecasts based on three classes of data (type of offence, place where the offence was committed, and the date/time of its commission). These forecasts are then used to identify, on a Web interface based on Google Maps, the high-risk areas in certain time periods. These results should therefore make it possible to optimise the distribution of human resources and equipment by directing police officers to the areas where the risk is higher.[10] Other examples of such predictive systems, often specialised in fighting specific crimes, can also be found in Europe. Among the best known is KeyCrime, developed by the Milan Police and used to predict robberies in the metropolitan area,[11] and XLAW,[12] developed by the Naples Police and applied by law enforcement agencies in several Italian regions to predict thefts and robberies.

The main advantages of these more "conventional" systems are twofold. Firstly, they contribute to better management of law enforcement know-how in a specific geographical area, freeing its preservation from the physical presence and expertise of individual agents. Secondly, these systems can improve investigative performance under conditions of limited human resources, while allowing for a more efficient allocation of investigative resources. Of course, the use of such systems can also help to reinforce possible discrimination, as by encouraging the police to oversee certain geographical areas and making it more likely that the people living there will be subject to stop-and-frisk practices. As will be discussed,[13] however, these

---

[9] W.L. Perry et al, *op. cit*.

[10] <https://www.predpol.com/law-enforcement/#predPolicing>.

[11] <https://www.emmeviemme.com>.

[12] E. Lombardo, *Sicurezza 4P, Lo Studio alla base del software XLAW per prevedere e pervenire i crimini* (Me Publisher, 2019).

[13] See § 3.2.

discriminatory effects are not necessarily created by the use of A/AI systems. In fact, algorithms and artificial intelligence are limited, in that they "only" perpetuate criticalities already found when such activities are carried out by human beings. In this sense, this application of A/AI technologies can be considered less problematic for the protection of fundamental rights, at least when compared to uses that directly apply to individuals.

2.2 Individual Risk Assessment

A much less conventional approach is to calculate the probability of individual risk. By accessing huge amounts of data, even if not necessarily already available to law enforcement, these systems correlate statistical risk factors with specific individuals, thanks to mathematical models and machine learning techniques.

Perhaps the best known of these predictive systems is COMPAS, developed by Northpoint Inc. (now Equivant), a private company based in California, and currently adopted in several US states to calculate the rate of recidivism, as when issuing decisions on alternative measures or suspended sentences.[14]

COMPAS predictions are based on information defined as "static" (e.g., criminal records) and on "limited use" of certain "dynamic" variables (e.g., drug abuse).[15] However, due to intellectual property rights on the software, no further details are available on how these variables affect the evaluation carried out by the system. What is known is that part of the information used by COMPAS is derived from the answers the person under evaluation (mostly, the suspect or the convicted) gives in response to a questionnaire of 137 questions, covering different aspects of the individual's personality and history.[16] By comparing data from similar stories, COMPAS assesses the risk of individual recidivism on the basis of three risk factors ("pre-trial," "general," and related to violent behaviour), assigning a numerical score to the individual. Therefore, COMPAS does not only make assessments based on elements specific to the individual but also extracts individual assessments from group data.

Another tool for individual risk assessment, especially in the pretrial phase, is the Public Safety Assessment (PSA), developed by the Laura and John Arnold Foundation and currently used in dozens of jurisdictions in the United States and in

---

[14] *State of Wisconsin v. Loomis*, 881 N.W.2d 749 (Wis. 2016), 36.

[15] <http://equivant.wpengine.com/classification/>. Cf. A. Widgery, *National Conference of State Legislatures, Trends in Pretrial Release: State legislation* (March 2015), at <https://comm.ncsl.org/productfiles/98120201/NCSL-Pretrial-Trends-Report.pdf>.

[16] J. Skeem, J. Eno Louden, "Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)," 2007, at <https://ucicorrections.seweb.uci.edu/files/2013/06/CDCR-Skeem-EnoLouden-COMPASeval-SECONDREVISION-final-Dec-28-07.pdf>.

some of the largest cities in the country, such as Phoenix, Chicago, and Houston.[17] The PSA makes two types of predictions, calculating (i) the risk that the person will not appear in court at the hearing and (ii) the risk of recidivism in the event of early release (with particular attention to violent crimes). There are nine factors considered by the system; as is known, they include age, pending proceedings, and criminal records. The risk is calculated on a scale of 1 to 6, where the highest scores indicate a higher level of risk.[18] Taking their cue from the numerous criticisms of COMPAS, the creators of this system decided to make information about its functioning public and, in particular, to reveal the different weight of each of these nine factors in the final calculation.[19]

However, the use of individualizing predictive tools is not a prerogative of the United States. The Harm Assessment Risk Tool (HART), developed by the Durham Police and the University of Cambridge, for example, makes predictions based on 33 different metrics, including the offender's criminal record, age, and postcode. Like COMPAS, HART also classifies individuals into high, moderate, or low-risk groups. The parameters used by HART have been made at least partly accessible to the public. In this way, it has been possible to identify a number of relevant criticalities—an operation that, for reasons of intellectual property, is not possible with COMPAS. The available information, for example, shows that the HART software is designed to promote false positives over false negatives, which means that it is more likely that a low-risk individual is wrongly classified as a high-risk person rather than the other way around.[20]

3. Automated risk assessment and fair trial rights: what limits?

The use of A/AI technologies to formulate individual predictions in criminal matters raises questions about various aspects of the right to a fair trial.

This is the case, for instance, with the right of access to justice. Indeed, A/AI systems assign scores to individuals but do so without processing data referring to each specific individual case: They analyse group data and then allocate the individual within a potential average classification. In essence, they produce a probabilistic result, which may be accurate, but does not coincide with the idea of individual

---

[17] E.g. Arizona, Kentucky, and New Jersey, <https://www.psapretrial.org/about>.

[18] Cf. Public Safety Assessment, *Risk Factors and Formula*, at <https://www.psapretrial.org/about/factors>.

[19] Critical about the equity of PSA, K. Patrick, "Arnold Foundation to Roll out Pretrial Risk Assessment Tool Nationwide," 3.09.2018, <https://www.insidesources.com/arnold-foundation-to-roll-out-pretrial-risk-assessment-tool-nationwide/>.

[20] M. Oswald, J. Grace, S. Urwin, G. C. Barnes, "Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality," *Information & Communications Technology Law* (2018), 27:2, 223–50, 236.

assessment of the case that lies behind Article 6 of the European Convention on Human Rights (ECHR) and Article 47 of the Charter of Fundamental Rights of the European Union (CFR). Both these provisions, requiring that "everyone is entitled to a fair […] hearing" before a tribunal, indeed logically assume the case to be assessed with regard to the specific position of the accused. Samples of comparable positions referring to different subjects may be considered by the court in its reasoning, but they cannot substitute the need for the judges to justify the sentence in relation to the specific and unique circumstances pertaining to the defendant(s). Violations of this basic rules, therefore, can hardly be considered minor and suitable to be compensated by the "overall" approach generally applied (but on different profiles) by the ECtHR in assessing fairness.[21]

The use of predictive systems may also undermine the presumption of innocence, for instance when a decision to apply a penalty or a pretrial measure is solely or in large part based on algorithms that attribute a higher risk status in case of doubt (as we saw with HART). If the presumptions (of risk) established by A/AI systems results *de facto* irrebuttable for the defendant,[22] this could represent another violation of Articles 6 ECHR and 48 CFR. According to the ECtHR settled case-law, indeed, presumptions of fact or of law operate in every criminal-law system and are not prohibited as such by the Convention.[23] Nonetheless, they should be confined to "reasonable limits", that effectively allow the defendant to overcome them.[24]

Automated assessment tools like COMPAS, in which part of the data used for calculations derives from information provided by the person to be evaluated, may be also problematic when it comes to the right to remain silent. Critical, in particular, is the uncertainty about the right of individuals to decline to fill out the questionnaire, on the basis that its results may be used to support their conviction or to apply a more severe penalty.[25]

In this contribution, for the reasons illustrated below, we focus on the right to an effective remedy.

---

[21] For instance, the Court has found a violation of Article 6§1 where an assize court refuses to put distinct questions in respect of each defendant as to the existence of aggravating circumstances, thereby denying the jury the possibility of determining the applicant's individual criminal responsibility (2 June 2005, *Goktepe v. Belgium*, appl. no. 50372/99, §28).

[22] In light of the lack of transparency on the functioning of the system and on the dataset used to produce the outcome, as well as due to the current incapacity of A/AI systems to state reasons for their decisions, as it is further argued in §§ 4.1 and 4.2.

[23] Cf., *e.g.*, ECtHR, 19 October 2004, *Falk v. the Netherlands*, appl. no. 66273/01.

[24] Cfr., e.g., 7 October 1988, *Salabiaku v France*, appl. no. 10519/83, § 28; 30 March 2004, *Radio France and others v France*, appl. no. 53984/00, §§ 23-24.

[25] C. Deskus, "Fifth Amendment Limitations on Criminal Algorithmic Decision-Making," *NYUJ Legis. Pubs & Pubs. Pol'y* 21 (2018), 237, 250.

Previously, however, it shall be stressed out that these and other[26] fair trial violations easily emerge where recourse to automated risk assessment is made compulsory by law, as it is in many US jurisdictions.[27] No less complex, though, are situations where A/AI systems developed for a certain purpose end up being used in criminal proceedings for other purposes as well.

The best-known case in this regard is certainly *Loomis*, decided in 2016 by the Wisconsin Supreme Court. In this case, object of much legal scholarship,[28] the Court was called on to examine the many criticalities of COMPAS, originally developed for probation support, in its application to the sentencing phase. First, the software had been publicly accused of being discriminatory, especially against African Americans.[29] Second, the system had been validated in some jurisdictions, but not in Wisconsin, so it was not clear whether its metrics were accurate for that target population as well.

The second criticism is particularly relevant in light of the well-known *Daubert* decision, where the US Supreme Court (USSC) held that expert evidence can be admitted only if scientific methods it is based on are proven to be reliable.[30] The USSC has not yet ruled on whether the *Daubert* criteria are applicable to A/AI systems or to sentencing, and the latter is generally excluded at the state level.

However, a line of reasoning similar to *Daubert* has been applied by the Supreme Court of the District of Columbia in a case concerning SAVRY, an

---

[26] Such as blurring of the line between preventive and post-factum inquiries, or for equality of arms, see S. Quattrocolo, "Quesiti nuovi e soluzioni antiche? Consolidated regulatory paradigms vs. risks and fears of 'predictive' digital justice," 4 *Cass. pen*. (2019), 1748 ff.

[27] E.g. for probation, early release, or bail. C. Doyle, C. Bains, B. Hopkins, *Bail Reform: A Guide for State and Local Policymakers* (Criminal Justice Policy Program, Harvard Law School, February 2019).

[28] State of Wisconsin v. Loomis, 881 N.W.2d 749 (Wis. 2016); cf. Recent cases, 130 Harv. L. Rev., 2017, 1530ff.; Eric L. Loomis, Petitioner v. State of Wisconsin, on Petition for a Writ of Certiorari to the Supreme Court of Wisconsin: Brief for the United States as Amicus Curiae, at <https://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf>; I. De Miguel Beriain, "Does the Use of Risk Assessments in Sentences Respect the Right to Due Process? A Critical Analysis of the Wisconsin v. Loomis ruling," 17 Law, Probability and Risk (2018), 45–53; S. Quattrocolo, op. cit.; M. Gialuz, "Quando la giustizia penale incontra l'intelligenza artificiale: luci e ombre dei rischi assessment tools tra Stati uniti ed Europa," Dir. pen. cont., 29.05.2019.

[29] Cf. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks," ProPublica, 23.05.2016, at <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

[30] Assessing whether the theory has been tested; is validated by *peer review* literature in the scientific field; has an assessed potential error rate; and is supported by standards that control its functioning. See *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).

algorithmic risk assessment tool used to predict violent behaviour in minors.[31] While not explicitly citing or rejecting the *Daubert* application, the DC court found the SAVRY assessment to be inadmissible, considering it unproven that the specific application of the software (not the software itself) had been carried out according to scientifically sound criteria.

In *Loomis*, these arguments were not deemed sufficient to find the COMPAS assessment inadmissible. Ignoring the allegations of discriminatory effects, the Wisconsin court arrived at its decision by drawing up a test based on whether the contested evidence was decisive or exclusive—a test in a sense similar to that developed by the ECtHR, in other contexts and for other purposes.[32] In the Wisconsin court's opinion, no due process violation can be recognised if A/AI systems have been correctly applied and if the automated assessment is supported by other elements.[33]

The reasoning developed in *Loomis* is not an isolated case overseas. As early as 2010, in the *Malenchik* decision, for example, the Indiana Court of Appeals supported the use of automated risk assessments as a basis for finding aggravating circumstances during sentencing. In this case, too, the defendant complained that the use of this system had been previously disapproved[34] and that the algorithm was unreliable and discriminatory because its calculation models had not been recognized as scientifically reliable in Indiana. However, the Court considered such use legitimate, since it did not completely *replace* the judge's discretion and was corroborated by other independent elements.[35]

In general terms, then, US courts seem to agree on a wide use of A/AI systems—originally developed for preventive or enforcement purposes—even in sentencing, so long the decision is not based solely or exclusively on automated risk assessments. Due to the lack of any specific case-law in Europe on the matter, this jurisprudence is here examined to assess whether it can be found effective in protecting the fundamental rights of the accused in our continent as well.

4. (Partly) Automated Decisions and the Right to an Effective Remedy

---

[31] Structured Assessment of Violence Risk in Youth (SAVRY), see Supreme Court of the District of Columbia, Justice Okun, 25.03.2018, as reported by S. Quattrocolo, *op. cit.*

[32] E.g., 8 February 1996, *Murray v United Kingdom*, App No 18731/91; 15 December 2011, *Al-Khawaja and Tahery v United Kingdom*, appl. no. 26766/05 and o.; 27 June 2017, *Chiper v Romania*, appl. no. 22036/10.

[33] In the present case, three read-in charges. The decision became final, see *Certiorari denied*, 137 S. Ct. 2290 (2017).

[34] *Rhodes v. State*, 896 N.E.2d 1193, 1195 (Ind. Ct. App. 2008).

[35] Level of Service Inventory-Revised (LSI-R) and Substance Abuse Substance Subtle Screening Inventory (SASSI). Cf. *Malenchik v. State*, 928 N.E.2d 564, 574 (Ind. 2010).

The right to an effective remedy, perhaps more than any other right, risks being dramatically jeopardized by automated assessments.

Provided for by Article 13 ECHR and Article 47 CFR, this right is at the same time one of the most important and least defined aspects of the notion of due process.[36]

An in-depth and exhaustive analysis of the concept of effective remedy, still highly debated, would go far beyond the scope of the present work. Nonetheless, few fundamental elements referring to this right may be identified in the jurisprudence of the European Courts. In the case-law of the Strasbourg Court, a remedy can be regarded as effective only if available both on the books and in action, that is, only if it can prevent an alleged infringement from persisting or can at least provide an adequate response for past infringements.[37] Thus, it is not sufficient for a remedy to be established in national law: Its effectiveness should be concretely assessed, as by considering the time it takes for the corrective action to be taken or the applicant's effective ability to activate the remedy in light of the specific circumstances of the case.

According to the ECtHR, it is not strictly necessary for the appeal to be lodged with a judicial authority; however, the empowered authority shall comply with the independence and impartiality requirements set forth in Article 6(1) ECHR. On this point, a difference may be observed between Article 13 ECHR and its analogue in Union law. Indeed, Article 47(1) CFR expressly requires that any violation of the fundamental rights enshrined in the Charter itself be effectively challengeable before a court.[38]

Most importantly, in the criminal matter, both the courts in Strasbourg and in Luxemburg require that decisions imposing a punitive measure shall be granted a full

---

[36] Also defined as the Convention's "darkest" provision. See Judges Matscher and Pinheiro Farinha in 2 August 1984, *Malone v. United Kingdom*, appl. no. 8691/79. On the vagueness of the notion of effective remedey also in EU law, see A. Soo, "Article 12 of the Directive 2013/48/eu: A Starting Point for Discussion on a Common Understanding of the Criteria for Effective Remedies of Violation of the Right to Counsel", 25 *EJCCLCJ* (2017), 31-51.

[37] See ECtHR, 26 December 2000, *Kudła v. Poland*, appl. no. 30210/96, §§ 157–58.

[38] The CJEU already established this principle before the entry into force of the Charter, cf. 15 May 1986, *Marguerite Johnston v Chief Constable*, Case 222/84, ECLI:EU:C:1986:206, 1651; 15 October 1987, *Union nationale v Georges Heylens and o*, Case 222/86, ECLI:EU:C:1987:442, 4097; 3 December 1992, *Oleificio Borelli SpA v Commission*, Case C-97/91, ECLI:EU:C:1992:491. As noted by A. Soo, *op. cit.*, 46 "It will be very interesting to observe how the ECJ will develop its approach to 'effective remedy'". To date, at least two requests for preliminary ruling to clarify the meaning of 'effective remedy' are pending before the Court of Justice, although not referring to the criminal matter, cf. joined cases C-67/20, C-68/20 and C-69/20 (Dublin III Regulation 604/2013), and case C-831/219 (Unfair Terms Directive 93/13).

judicial review. Thus, it is necessary to identify at least one authority with the power to rule both on questions of fact and of law.[39]

The following analysis on the right to an effective remedy will take its lead from these elements as defined by the European courts.

The right to an effective remedy is regularly reaffirmed in EU secondary legislation on criminal procedure,[40] and on the protection of personal data.[41] None of these statutory acts, however, provide a detailed definition of how a remedy needs to be structured to be really effective.

In light of this intrinsic vagueness, the right at stake, already problematic in many contexts, becomes exceptionally critical when decisions imposing punitive measures on the accused are (even partly) automated.

There are several circumstances that determine this criticality. Firstly, it is difficult for the accused to assert their right to an effective remedy without access to all the necessary information grounding the decision. Secondly, the duty to state reasons comes up against some extremely relevant technical obstacles with regard to automated assessments. Finally, and even apart from the above, when the decision follows nonhuman logic (at least partly), it is far from obvious how the notion of "effectiveness" should be defined. We will deal with this aspect in the final part of the paper.

---

[39] ECtHR, 23 October 1995, *Umlauft v Austria*, appl. no. 15527/89, § 37; 21 February 1984, *Öztürk v Germany*, appl. no. 8544/79 § 56; 27 September 2011, *Menarini Diagnostics S.R.L. v Italy*, appl. no. 43509/08, §§ 59-63-67; 23 October 1995, *Schmautzer v Austria*, appl. no. 15523/89, § 36; 23 October 1995, *Gradinger v Austria*, appl. no. 15963/90, § 44.

[40] Directive 2010/64/EU of the European Parliament and of the Council of 20 October 2010 on the right to interpretation and translation in criminal proceedings, OJ L 280, 26.10.2010, p. 1–7; *Directive* 2012/13/EU of the European Parliament and of the Council of 22 May 2012 on the right to information in criminal proceedings, OJ L 142, 1.6.2012, p. 1–10; Directive 2013/48/EU of the European Parliament and of the Council of 22 October 2013 on the right of access to a lawyer in criminal proceedings and in European arrest warrant proceedings, and on the right to have a third party informed upon deprivation of liberty and to communicate with third persons and with consular authorities while deprived of liberty, OJ L 294, 6.11.2013, p. 1–12; Directive (EU) 2016/343 of the European Parliament and of the Council of 9 March 2016 on the strengthening of certain aspects of the presumption of innocence and of the right to be present at the trial in criminal proceedings, OJ L 65, 11.3.2016, p. 1–11; Directive (EU) 2016/800 of the European Parliament and of the Council of 11 May 2016 on procedural safeguards for children who are suspects or accused persons in criminal proceedings, OJ L 132, 21.5.2016, p. 1–20; Directive (EU) 2016/1919 of the European Parliament and of the Council of 26 October 2016 on legal aid for suspects and accused persons in criminal proceedings and for requested persons in European arrest warrant proceedings, OJ L 297, 4.11.2016, p. 1–8.

[41] Recital 104 of Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, OJ L 119, 4.5.2016, p. 89–131.

4.1 The Right to an Effective Remedy and Access to Information

First of all, in order to be able to effectively challenge an individual decision, it is necessary that the data subject has access to all the information relevant to the decision, and in particular to the datasets, the data processing methods, and the source code expressing the algorithms underlying the functioning of the system. Indeed, access to such information is in the interests not only of those directly affected by the decision but also of all the actors involved in the design, development, implementation, and use of A/AI systems in criminal justice, including programmers and judges; and, more generally, the information is relevant to public opinion as well.[42]

However, in order to make access to relevant information truly effective, a certain degree of transparency in the decision-making process is necessary. This requirement is also expressed in Recital 38 of Directive 2016/680, stating that "in any case, such [automated] processing should be subject to suitable safeguards, including the provision of specific information to the data subject and the right to obtain human intervention, in particular to express his or her point of view, to obtain an explanation of the decision reached after such assessment or to challenge the decision."[43]

However, A/AI systems often present significant challenges for transparency: information on the dataset is usually not available to the parties or the judge using the system, and a similar consideration may be made with regard to the information on the data processing methods and algorithms. In fact, in many cases they depend on the accessibility of the source code, the disclosure of which may be limited by intellectual property rights. Moreover, as discussed, in A/AI systems based on machine learning there are structural limits to the ability to provide information for reconstructing the system's functioning and the reasons for its decision.

In order to ensure a satisfactory level of transparency, a number of methods have been proposed, although it is still unclear how the information should be made available in practice and which elements should be included.

---

[42] IEEE, *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*, version 1, IEEE Standards Assoc., 2016, at <https://standards.ieee.org/news/2019/ieee-ead1e.html>.

[43] Language nearly identical to Recital 71 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation - GDPR), OJ L 119, 4.5.2016, p. 1–88.

A first option, suggested in the Ethical Charter of the Council of Europe,[44] is full technical transparency, i.e., the disclosure of both the source code and the accompanying documentation. However, as mentioned, when the system is developed by private entities, as in the case of COMPAS, access to the source code may be limited for intellectual property reasons and in virtue of the need to protect trade and industrial secrets.

Even in cases where it is possible to access the source code, however, this may prove to be only a partial solution to the problem of transparency, especially with a view to ensuring an effective remedy. In fact, not only is the source code of A/AI systems usually incomprehensible to nonexperts, but even programmers often find it difficult to understand how these systems work and to predict their results solely by inspecting the source code.

These limitations have been clearly identified in several court decisions across Europe, particularly in the field of school education, where A/AI systems are commonly used in classifying and selecting students and in assigning teachers to plexuses.

In France, for example, the *Admission Post Bac* algorithmic system has long been used in the procedure for enrolling students at university. In 2016, the *Commission d'Accès aux Documents Administratifs* declared itself in favour of giving access to this platform's source code. Although the source code was made available, the experts were not able to reconstruct the system's complete reasoning. This was determined not only by the algorithm expressed in the source code, but also by the different input data coming from a set of databases external to the system. Without disclosure of these data, and without information on the structure of the tables and the description of the fields used in the databases containing the data, mere disclosure of the source code was not sufficient to ensure an effective remedy.[45] Similar considerations can be made with regard to some Italian cases concerning the assignment of teachers to high schools. In 2017, for example, the Lazio Regional Administrative Court established that the Ministry of Education is obliged to issue a copy of the source code of the software used in this procedure. However, even in this case, no mention was made of the disclosure of the *input data*, the structure of the table data, or the description of the fields used in the databases linked to the system.[46]

A second option for solving the transparency problems is to also disseminate top-level information on the logic of automated decision-making, possibly in natural language, i.e., in a language understandable even to lay users.

---

[44] Council of Europe, CEPEJ, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*, adopted on 3–4.12.2018, at <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>, 11.

[45] Decision of the Commission d'Accès aux Documents Administratifs of 23.06.2016, no. 20161990.

[46] TAR Lazio, Sezione Terza bis, Decision no. 03769/2017 (hearing of 14.02.2017).

Indeed, as the Article 29 Working Party has pointed out, "complexity is no excuse for failing to provide information."[47] This is the approach taken by the GDPR at Article 13(2)(f), requiring that when the personal data used in an automated decision-making process are collected from the data subject, "significant information [be provided] about the logic used" by the system. The same provisions are reiterated in Article 14(2)(g) in relation to data not obtained by the data subject.

To that end, the information to be disclosed should at least include information on the data that served as input for the automated decision; information on the list of factors that influenced the decision; information on the their relative weight in the decision; and a reasonable explanation (possibly in textual form) of the reasons why a certain decision was made.[48]

The criterion last mentioned, however, is particularly critical to assess. Some authors wonder whether the explanation should provide complete information about *all* the models and variables taken into account by the system (a *model*-centric explanation), or only about those models that are relevant to the specific case under consideration (a *subject*-centric explanation).[49] Furthermore, it remains uncertain whether the explanation should also include third parties' personal data used for the decision. This option is subject to limitations by the GDPR itself. However, in some cases it may be necessary to assess the fairness of the decision made in relation to the situation of different persons in comparable situations.

Another critical point concerns the extent to which the information should be detailed with regard to each input, as well as its weight in the decision. The disclosure of a full explanation of the model could in fact have side effects. For example, in a system for automated detection of tax evasion, the disclosure of the risk thresholds used by the system may encourage the adoption of "strategic behaviours" by individuals filing their tax returns.

For these reasons, rather than requiring transparency, several experts suggest focusing on *procedural regularity*, i.e., the adoption of specific techniques that demonstrate the system's ability to meet certain standards of fairness even in

---

[47] Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679* (wp251rev.01), of 3.10.2017, <https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=49826>, note 40 and p. 25. The Article 29 Working Party (Art. 29 WP) was an advisory body providing recommendations and promoting the consistent application of the Data Protection Directive, whose composition and purpose was set out in Article 29 of the Data Protection Directive. In 2018, with the entry into application of the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679, it has been replaced by the European Data Protection Board (https://dpb.europa.eu).

[48] M. Brkan, "Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond," 27 *International Journal of Law and Information Technology*, 2 (2019), 91–121, 113.

[49] L. Edwards, M. Veale, "Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For," *Duke L. & Tech. Rev*. (2017), 16:18.

automated decisions, without revealing which key attributes are used in the decisions, or the details of the underlying algorithmic processes.[50] Other scholars, however, stress that this procedural regularity only ensures that decisions are based on the same decision-making policy, that the policy has been determined before knowing the input data, and that the results can be reproduced. It therefore only considers the aggregate procedural regularity of all cases, ensuring that they are decided according to the same rules. But not even the concept of procedural regularity explains why the algorithm has reached that specific individual decision against that person.[51]

4.2 The Right to an Effective Remedy and the Duty to State Reasons

The duty to state reasons, especially in criminal or punitive matters, is a prerequisite for exercising the right to an effective remedy.[52] However, this assumption risks being circumvented when decisions are based on automated evaluations, particularly when these decisions results from A/AI systems based on machine learning.

In fact, the impossibility of reconstructing the internal functioning of A/AI systems often translates into a fideistic approach to the result provided, that is, into believing that the decision is justified by the very fact that, with a certain statistical precision, the decision was made by the system itself. This approach has been termed *data fundamentalism*, i.e., the tendency to believe that the analysis carried out using *data mining* techniques on large datasets always provides an objective view of reality, ignoring the fact that the correlations identified by the algorithm, and on which the decisions are based, do not necessarily imply a causal link.[53] As Kroll et al. observe, the analysis and decisions made by computers often enjoy an undeserved assumption of fairness or objectivity, even if the design and implementation of automated decision-making systems are exposed to critical issues that can lead to systematically

---

[50] "Computer systems can be designed to prove the oversight authorities and the public that decisions were made under an announced set of rules consistently applied in each case, a condition we call *procedural regularity*". A. Kroll, S. Barocas, E. Felten, J. R. Reidenberg, D. G. Robinson, H. Yu, "Accountable Algorithms," 165 *U. Pa. L. Rev.* (2016), 637.

[51] M. Brkan, *op. cit*.

[52] According to established ECtHR case-law, the duty to state reasons reflects "a principle linked to the proper administration of justice […]Without requiring a detailed answer to every argument advanced by the complainant, this obligation presupposes that parties to judicial proceedings can expect to receive a specific and explicit reply to the arguments which are decisive for the outcome of those proceedings" (11 July 2017, *Moreira Ferreira v Portugal*, appl. no. 19867/12, § 84). Courts must "indicate with sufficient clarity the grounds on which they based their decision. It is this, inter alia, which makes it possible for the accused to exercise usefully the rights of appeal available to him" (16 December 1992, *Hadjianastassiou v. Greece*, appl. no. 12945/87, § 33).

[53] K. Crawford, "The Hidden Biases in Big Data," *Harvard Business Review Blog Network*, 1.04.2013; B. Prietl, "Big data: Inequality by design?" *Weizenbaum Conference* (2019), DEU, 10.

erroneous and biased decisions.[54] For these reasons (lack of transparency), it will also often be very difficult to identify potential biases that influence a system's decision-making.

As regards machine learning systems in particular, the main causes of bias, which can result in discriminatory effects, are primarily related to problems with the dataset, and in particular to (i) the use of a dataset containing data reflecting an implicit or explicit bias that from the outset is built into in the decisions on which the system is trained;[55] and (ii) the use of a dataset containing data offering a statistically distorted picture of certain groups in relation to the overall population.[56] Moreover, even datasets without errors or initial biases can lead to discriminatory decisions, this owing to the inability of machine learning systems to distinguish between mere correlation and causality, as well as to the effects of the model's self-reinforcement based on new data incorporated into the dataset.[57] Another cause of discrimination, specific to the criminal sector, is that most (if not all) of the A/AI systems used in this area only refer to a limited number of crimes, often referred to as "street crimes." This contributes to increasing the perception that some offenders are particularly dangerous, while ignoring that other types of crimes are also accompanied by high rates of recidivism, such as white collar crimes. Finally, since most predictive software is developed and/or owned by private for-profit companies (as in the case of COMPAS), the determination of the dataset's content, or of the selection process applied by the dataset algorithm, could also lead to discriminatory decisions on the basis of undeclared potential conflicts of interest (e.g., of commercial nature).

From a different perspective, however, some scholars argue that, far from amplifying discriminatory effects, the increase in A/AI systems, properly used, could in the future in fact correct and limit the harmful effects of cognitive human biases, especially those of judges. For example, with regard to decisions on setting bail, a recent study has shown that an algorithm designed to predict the risk of failure to appear in court has obtained more equitable results than human judges, because, unlike the latter, in making the decision it was not influenced by the bias related to

---

[54] A. Kroll, S. Barocas, E. Felten, J. R. Reidenberg, D. G. Robinson, H. Yu, "Accountable Algorithms," 165 *U. Pa. L. Rev.* (2016), 633.

[55] E.g. if a system is trained on human decisions that contain some racial prejudice, the system will reproduce discriminatory results "inherited" from human decisions.

[56] E.g. a predictive police system trained on a dataset that overrepresents the incidence of crimes in certain ethnic groups. Law enforcement would be directed by the algorithm to check more people in these groups, with the result that, statistically, more offenses will be discovered within it. When data on new crimes are added to the dataset, overrepresentation will increase, reinforcing the discriminatory effect. See K. Miller, "Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm," 19 *J. Tech. L. & Pol'y* (2014), 105.

[57] B.E. Harcourt, "Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age', 94 *University of Chicago Public Law and Legal Theory Working Paper* (2005), 36–37.

the seriousness of the crime (current accusation bias).[58] Positive effects for the defendants were also found in another study evaluating the use of the PSA system in Lucas County, Ohio, where the software was adopted in 2015. In this case, thanks to the PSA, there has been an increase in the number of persons released without recourse to bail, as well as a significant reduction in the number of crimes committed by defendants awaiting trial who are not subject to precautionary measures.[59]

A precaution usually taken to prevent or mitigate the risk of discriminatory effects consists in excluding or removing sensitive data (data about racial or ethnic origin, political opinions, religious beliefs, health, sexual orientation, etc.) from the dataset.[60] However, A/AI systems may also be used to extract sensitive data from the processing of nonsensitive personal data. For example, in a famous case, an A/AI system the US retailer Target used to analyse customer purchases was able to assign to each customer a "pregnancy prediction" score, with an estimate of the date of delivery, based only on an analysis of the pattern of purchases of certain products, coupled with some additional demographic information.[61]

Some recent contributions have suggested other possible actions by which to manage and limit the risk of bias and the resulting discriminatory effects, such as (a) ensuring and tracing the origin of the data used by the system (and, where appropriate, certifying its sources), as well as their quality and coverage, and making sure that they have not been altered before being used by the machine learning system, so as to make the entire data lifecycle traceable; (b) providing information on the data processing methods, possibly by means of an independent audit, where direct access to the source code is not possible;[62] and (c) providing the individual affected by a decision made using a black box system with a set of counterfactual explanations, i.e., information describing the smallest changes to the inputs of the system that, by hypothesis, would have led to a different and desirable result for the person concerned, without having to explain the system's internal logic. In this way, knowing which external factors and variables have contributed to an automated evaluation, the person who is subject to the decision would be able to challenge it and, in particular, to obtain evidence of

---

[58] R. Sunstein, "Algorithms, Correcting Biases," 86 *Oxford Business Law Blog. Social Research: An International Quarterly*, 2 (2019), 499–511; J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, "Human Decisions and Machine Predictions," 133 *The Quarterly Journal of Economics*, 1 (2018), 237–93.

[59] J. Tashea, "Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions," 1.03.2017, at <www.abajournal.com>.

[60] See "special categories of personal data" in Art. 9 GDPR.

[61] L. Floridi, The Fourth Revolution: How the Infosphere Is Reshaping Human Reality (Oxford: OUP, 2014), 16.

[62] CEPEJ, *op. cit.*, 11.

possible discrimination when the evaluation is determined by sensitive data (e.g. race or ethnicity).[63]

4.3 What Remedy Is Effective Enough?

Even in cases where no discrimination or transparency issue is at stake, several critical issues remain in relation to the same definition of "effective remedy". Indeed, remedies presently available in the EU do not really seem effective when it comes to automated or partly automated decision-making.

In European countries, where A/AI systems are still much less employed than overseas, there is little in the way of specific tools for reviewing algorithmic decisions in criminal matters.[64] In the EU law, however, the matter has been addressed at least since 2016. Article 11(1) of Directive 2016/680, in particular, affirms that "a decision based solely on automated processing, including profiling, which produces an adverse legal effect [shall] be prohibited unless authorised by [...] law [...] which provides appropriate safeguards [...], at least the right to obtain human intervention."

This right to human supervision has so far generated an interesting jurisprudence in the Member States, in areas other than criminal law.[65] That is the case, for example, in Italy, where in 2018 the Lazio Regional Administrative Court held that a discretionary administrative procedure (the allocation of school staff) cannot be fully entrusted to an algorithm.[66]

Ex post human supervision, however, is a "remedy" that presents several critical issues, especially in criminal matters. First, drawing a clear distinction between fully automated and semiautomated decision-making may sound very logical and appealing. The idea was also expressed by the Article 29 Working Party, according to which "if a human being reviews the outcome of the automated process and takes into account other factors in making the final decision, that decision will not be 'solely'" automated.[67] In practice, however, the boundaries between these two models are blurry to human eyes.[68]

---

[63] S. Wachter, B. Mittelstadt, C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," 31 *Harv. JL & Tech.* (2017), 853.

[64] Different in other areas of law, e.g. German tax law where administrative acts may be automatically adopted, if allowed by law and no discretionary assessment is required. See Gesetz zur Modernisierung des Besteuerungsverfahrens vom 18.07.2016 (BGBl. I S. 1679), § 35a.

[65] The same principle, in relation to personal data protection, can be found in Article 22 GDPR.

[66] TAR Lazio, Sezione Terza Bis, decision no. 09230/2018 (hearings of 26.06.2018 and 11.07.2018).

[67] Article 29 Data Protection Working Party, *op. cit.*

[68] Council of Europe, *Study on the Human Rights Dimensions of Automated Data Processing Techniques (in particular algorithms) and Possible Regulatory Implications,* 6.10.2017.

In fact, it is questionable whether the human beings charged with supervising automated decisions are in a position to do so—so much so that A/AI systems are themselves not yet technically capable of providing intelligible explanations of their own reasoning. And not just so: as claimed by an authoritative American scholar, AI is fundamentally alien to human intelligence, and in fact the purpose of such systems is often precisely to learn to do or see things in ways impossible for human beings.[69] This is all the more true once we consider that human capacity for judgement is claimed to be inferior to that of mathematical models when it comes to prognostic evaluations.[70] Even where humans formally retain control over the final decision, the possibility of effectively contesting its merits remains at best a remote hypothesis.

The question of who is effectively responsible for making decisions when A/AI systems are involved has already been raised in a number of contexts that, like criminal law, stand apart in virtue of how vital the interests at stake are, and how essential it is that decisions be timely. An example is the medical area, where A/AI systems are used to support physicians by generating diagnoses and treatments with a level of accuracy that is often greater than that of the corresponding human assessments.[71] Here it is unlikely that a human supervisor, even a highly trained one, can actually review the merits of an automated evaluation. Apart from catching obvious errors, physicians will basically have only two alternatives, especially when pressed for time: they can trust the automated assessment, because they place trust in the A/AI system that generated it; or they can decide *not* to trust the system, and hence its result. Indeed, the lack of substantial elements on which basis to challenge AI predictions ends up reducing the supervisory task to deciding whether or not to rely on the AI system.[72]

A similar challenge may also confront judges dealing with (partly) automated decisions in criminal cases, where the decision-making timeframe is generally less stringent,[73] and where few objective criteria are available for checking the accuracy of assessments (either machine-generated or human). Even in criminal matters human supervision over automated decisions is thus often reduced to a choice between trusting or not trusting the A/AI system. Against this background, the mere prospect

---

[69] A. D. Selbst, "Negligence and AI's Human Users" (11.03.2019), *Boston University Law Review*, forthcoming), available at SSRN, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350508>.

[70] A. Tversky, D. Kahneman, "Judgment under uncertainty: Heuristics and biases", *Science* 185(1974), 1124-1131; See also C. Deskus, *op. cit.*; and J. Millar, I. R. Kerr, *op. cit.*

[71] A case in point is IBM Watson Health. See <https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine>.

[72] A. D. Selbst, *op. cit.*

[73] But in closely related sectors J. Fergusson, "Twelve Seconds to Decide in Search of Excellence: Frontex and the Principle of 'Best Practice,'" Publications Office of the European Union (2014), 15, estimates that 12 seconds is about the time frame for Frontex to decide on the legality of individual applications for entry into EU territory.

of an *ex post* human intervention does not seem sufficient to guarantee an effective remedy.

5. All AI's Fault? "Human" Black Boxes and the Criminal Trial

In the lack of reasoning behind the functioning of A/AI systems lies, as mentioned, one of the biggest hurdles preventing a remedy from being effective. However, it would be wrong to assume that, in criminal proceedings, that problem comes up only when algorithms or AI are involved. Indeed, decision-making mechanisms that closely resemble the "oracle" model characterize also some totally "human" mechanisms in criminal proceedings and trials.

The jury is perhaps the most obvious example of such "human black boxes." In fact, in various legal systems, decisions on the merits of the charge are expressed in verdicts that do not state reasons. Similarly to automated decisions, then, even in trials by jury the accused may be slotted into one of two classes (either innocent or guilty) although neither the reasoning behind it nor any potential faults can be reconstructed with a reasonable degree of certainty. The parallel between these two decision-making models holds even considering jury selection procedures—a fundamental "human" tool with which to ensure some fairness in the face of the potential for discrimination—since specific measures against such risks can today be applied to A/AI systems as well.

Several examples of "human black boxes" may also be found within traditional, otherwise completely "explainable" procedural paradigms. These are cases where the adjudicating body is required to formulate a risk prognosis on the basis of "tacit" criteria which are, at least *de facto*, grounded in vague and not strictly legal notions, such as the judge's "intuition," "sense of justice," or "experience"[74] (e.g. in decisions concerning the application of measures alternative to detention).[75] It is true that human judges can (and do) state reasons for these diagnoses. However, it is rarely the case that such rationales contain—or even *can* contain—elements making it possible to objectively guarantee a higher level of fairness compared to that of A/AI decisions, unless that assessment is reduced to the mere enumeration of criminal records.[76] Moreover, the lack of transparency and the risk of discrimination are criticisms that—at a structural level—may also be directed at "human-only" decisions: human adjudication skills in formulating statistical and prognostic

---

[74] See W. Schulz, K. Dankert, "'Governance by Things' as a Challenge to Regulation by Law," 5(2) *Internet Policy Review* (2016).

[75] Cf. M. Caianiello, "Criminal Process Faced with the Challenges of Scientific and Technological Development," 27,4, *EJCCLCJ* (2019), 267–291.

[76] As in *three-strikes law* models, problematic in a system that prizes due process and the principle of the proportionality.

assessments are indeed often affected by bias, and are certainly influenced by the limits inherent in every human experience, however professional.

The margin of tolerance for potential mistakes is, however, commonly perceived as very different in the two cases. Although the possibility for a human judge to formulate risk prognoses based on vague and mostly unverifiable criteria—necessarily requiring a personal interpretation—as well as the power of the jury to issue "oracles," are sometimes criticized, they are usually accepted as legitimate. On the contrary, when such assessments are made by A/AI systems, their legitimacy tends mostly to draw opposition, especially in Europe. In this regard, some authors have pointed out that "we humans," while quite indulgent towards the weaknesses and failures of our own species, are much less tolerant of the possibility of failure in our machines. In other words, we probably expect far more from machines than we do from ourselves.[77]

Of course, human judges and juries are representatives of the (human) community, and therefore enjoy a degree of "political" endorsement difficult to accord to algorithms. And yet, as anticipated, a certain degree of democracy and transparency can be ensured even in the planning (and use) of A/AI systems, so that their results are aligned with the fundamental principles that ground our societies, foremost among them the principle of legality.

Moreover, in light of the technological development of AI, it is becoming increasingly difficult to justify a preference for human decision-makers on the basis that only human beings have the capacity not only to apply a rule but also to disapply it where necessary, looking for apparently unconventional solutions. AI does not actually belong to the traditional notion of "machine" as a mere enforcer of strictly programmed tasks; it is inherently characterized by a certain amount of "creativity," today still under development but which has already found quite some challenging applications, as in the creation of fashion collections.[78]

Even so, we instinctively continue to find more acceptable to ascribe adjudication skills to human judges, regardless of their unavoidable margins of error. It is beyond the scope of this contribution to do an in-depth analysis of the legal and political reasons behind this assumption or, perhaps even more importantly, of the psychological reasons behind it. Here, we do not want to suggest that this assumption should be abandoned, but rather that there may be an urgent need to rethink its foundation in our technologically advanced world. A/AI systems may serve as a stimulus for it.

---

[77] C. Allen, G. Varner, J. Zinser, "AMA: Artificial Moral Agents (Prolegomena to Any Future Artificial Moral Agent)," 12(3) *Journal of Experimental & Theoretical Artificial Intelligence* (2000), 251–61.

[78] *E.g.* Glitch brand (<https://glitch-ai.com/pages/about-us>), cf. J. Wood, "These Clothes Were Designed by Artificial Intelligence," *World Economic Forum* (2019), at <https://www.weforum.org/agenda/2019/07/these-clothes-were-designed-by-artificial-intelligence/>.

The growing debate on automated decisions, however, should also be welcomed as an opportunity to redirect our attention to "human" predictive assessments and their legitimacy in light of the fundamental rights of the accused. In the context of this debate, we may think of a new form of structured interaction aimed at enabling a shared decision-making model, that puts together the advantages of humans and technologies.[79] This model is sustainable though, only if humans are effectively able to retain the ability to oversee the A/AI activities. And this bring us back to the problem of ensuring an effective remedy.

6. Conclusion: Some Proposals towards a (Truly) Effective Remedy

In the light of the considerations so far made, it seems that the right to an effective remedy against automated or partially automated decisions should be based on the establishment of an integrated system of technical and legal guarantees, capable of preventing the use of A/AI in criminal justice from resulting in clear violations of the fundamental rights.

To this end, it is in the first place necessary that legal operators who deal with A/AI, and who will increasingly be doing so, have an adequate awareness of these systems' capabilities and limitations. It is not required that they become IT engineers, but it *is* essential that they know how to interact correctly with them, critically incorporating automated results into "human" evaluations. In other words, it is necessary to shift from an approach based on *data fundamentalism* to one based on *informed trust*.[80]

In the second place, rather than moving to a later stage the task of checking the accuracy of the A/AI system, entrusting persons who lack the necessary technical expertise, it would seem preferable to set up an *ex ante* certification mechanism that would allow the system to be validated with the participation or oversight of public authorities, as is already the practice in other sectors.[81] This certification should cover not only the functioning of A/AI system, but also the entire socio-technical system that encompasses technology, users (judges, prosecutors, law enforcement agencies, lawyers) and the legal and ethical rules governing such interaction.

In order to set up this certification mechanism, companies developing A/AI systems will be forced to produce information documenting the design approach,

---

[79] This model rests on the concept of a joint cognitive system. It has been observed that when humans and AI systems interact in working toward a goal, it would be better to describe humans and technology not as two interacting 'components', but as making up a joint cognitive system, where control is shared between the human cognitive system and the AI system. See E. Hollnagel, D. D. Woods, *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering* (CRC Press 2005).

[80] See IEEE, *op. cit.*, 220: "informed trust rests on a reasoned evaluation of clear and accurate information about the effectiveness of A/IS and the competence of their operators."

[81] IEEE, *op. cit*., 16.

development, quality, and scope of the dataset, operation, user training, etc. A requirement to produce such information would also constrain the development of the system itself: in order to meet the certification requirements, companies will be guided in the A/AI systems' design choices. This approach is already being followed in several "safety-critical" sectors (such as health care and aviation). In order to obtain such a certification, companies will have to produce solid evidence demonstrating that the system is suitable for its intended purpose. The certification should include rules by which to identify responsibility profiles, as by specifying what skills and knowledge are needed to use the system for a particular procedural outcome. An effective certification system should also establish risk classes based on the purpose and procedural stage at which the system will be used. Finally, certification should be complemented with a system of periodic audits, which may vary according to the risk class in which the system is used.

A model that could be adopted as a benchmark is that of Regulation 2017/745,[82] which sets out the necessary criteria through which a medical device is certified as a product complying with the appropriate safety and performance standards. Medical devices fall into four different classes, depending on the purpose of the device and the risks involved.[83] A different conformity-assessment procedure is defined for each class, requiring basic quality assessments for class I devices, up to full quality assurance for class III devices. While in the first case, the assessment of conformity can be done under the sole responsibility of the manufacturer, the complete quality-assessment procedure requires the involvement of a qualified body and a group of experts.[84]

As part of the certification process in criminal matters, it should be possible for critical data such as the source code to be inspected through an independent audit. This would strike the right balance between business (intellectual property, trade secrets) and public interest in controlling the use of this technology. It seems thus appropriate that audits be carried out in a uniform manner at national level, possibly by a public body capable of ensuring democratic control, such as a special parliamentary committee. Certification and validation can therefore also help to strengthen confidence in the use of such technologies in criminal matters, moving in the direction of what is currently referred to as "trustworthy" AI.[85]

Certification and validation could be considered as sufficient guarantees for the fairness of automated decisions made in contexts where risk assessment—the threat to individual defence rights—is relatively low. It could be the case, for example,

---

[82] Art. 10(9) Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, OJ L 117, 5.5.2017, p. 1–175.

[83] Idem, Art. 51.

[84] Idem, Articles 83(1) and 83(2). A risk-based approach is also suggested in: European Commission, White Paper on Artificial Intelligence: a European approach to excellence and trust, 2020, 17.

[85] European Commission, Ethics Guidelines for Trustworthy AI, 2019, 54.

with offences for which only financial penalties can be imposed. This would allow human resources to be more intensively devoted to detecting and fighting the most serious crimes.

In other cases, however, respect for fundamental rights and democratic control cannot be left exclusively to certification procedures.[86] These situations require the creation of innovative technical and legal solutions.

On the one hand, this means adopting "explicable" AI or XAI solutions. This expression refers to technical methods for explaining black box models, and in particular to those approaches in which machine learning methods are linked to symbolic or rule-based methods, so as to provide human understandable explanations including the complete logical-argumentative rationale for the decisions. To this end, according to some authors, it would be possible to provide an interpretable and transparent model that can replicate the black box's decision-making process, making it intelligible.[87] This approach is particularly interesting in criminal matters, where the ability to access all the factors examined, and the weight they each carry in the decision-making process, is more useful when it comes to protecting the right to defence than is a result that only expresses a numerical probability of a future event. The approaches based on explicable AI, however, are now in the development stage and have not yet reached a degree of maturity to be adopted in real-life contexts.

On the other hand, from a legal point of view, a first solution could be that of endorsing the *Loomis* perspective. This option, partly in line with the "overall" approach developed by the Strasbourg Court with regard to fair trial, does not, however, appear entirely satisfactory. As previously illustrated, it will always be difficult, if not impossible, for the defendant to contest the merit of an A/AI assessment, and this undercuts the ability to obtain a genuinely effective remedy against decisions based on such elements. As discussed, this gap can hardly be filled by the right to a subsequent "human" intervention. Article 11, Directive 2016/680 can thus be read as a statement of principle not to fully delegate decision-making power rather than as a provision that can achieve this result. Indeed, assessments produced by A/AI systems cannot be considered just as a different type of expert evidence, due to the fact that, for the reasons described above, judges don't have the skills to fully understand, and therefore review them.

In ensuring that remedies are truly effective, when the decision is (partly) automated, it seems more appropriate to change perspective.

For instance, the right to have assessments generated by one A/AI system reviewed by another automated system could be introduced in criminal proceedings. This approach, still unexplored in criminal justice, has long been applied in safety-critical areas, such as in aviation, where the use of *redundant* technologies is generally

---

[86] Idem, *op. cit*., 23.

[87] Cf. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," 51 *ACM Comput. Surv*., 5 (2018), 93:1–93:42.

considered the best way to reduce risks.[88] This means that the same information must be processed by a number of systems that are different but fulfil the same functions. Diversity can be achieved by adopting alternative approaches to algorithm development, employing different teams of programmers and selecting different components.[89]

Thus, for example, if a decision issued by a court of first instance is wholly or partly based on assessments made by system X, the defendant should have the right, on appeal, to have this assessment repeated by system Z. This approach would require that a range of certified and validated A/AI systems be made available in each judicial district (e.g., through the creation of a special register). Such systems should however also be designed and developed by different manufacturers. The range should be sufficiently broad that when an appeals court reviews the assessment produced in a trial court (or during a pretrial investigation), it can choose a system that is different from the one already employed.

Accessing a second automated assessment could indeed enable (human) judges to effectively apply criteria of (human) logic in making the comparison, and hence to implement (or try to implement) a truly effective remedy when requested by the accused.

The structural introduction of such review systems, which in a sense represent a technologically updated version of expert witnesses, could also help to reduce the disparity between defendants who can afford to examine automated assessments (and thus attempt to challenge them) and those who cannot. The risk of economic discrimination, present in all criminal trials, especially where it proves highly advantageous to recruit the best talent for one's defence, appears especially unfair in light of the current applications of A/AI systems, today exclusively used in connection with "street" offences, often involving economically disadvantaged defendants.

Certainly, an unregulated (and unreasoned) use of A/AI systems may diminish the ability to effectively exercise one's fundamental rights. However, algorithms and AI are proving to have great potential for transforming the entire decision-making dynamic in criminal cases, which so far has been mostly asymmetrically exploited. Perhaps the time has come to also put these technologies at the service of the defendant.

---

[88] Suggesting that different technologies be used to perform the same function, even when one is better than the others, H. Jones, "Common Cause Failures and Ultra Reliability," in 42nd International Conference on Environmental Systems (2012), 3602, <https://arc.aiaa.org/doi/abs/10.2514/6.2012-3602>.

[89] J. Downer, "When Failure is an Option: Redundancy, Reliability and Regulation in Complex Technical Systems," Discussion Paper no. 53, *Centre for Analysis of Risk and Regulation*, London School of Economics (2009).