The Compatibility of AI in Criminal System with the ECHR and ECtHR Jurisprudence

Nídia Andrade Moreira(✉)

Católica Research Centre for the Future of the Law, Porto, Portugal
nidiandrademoreira@gmail.com

Abstract. The admissibility of AI systems that focus on determining the measure of punishment must be analyzed in light of ECHR and ECtHR jurisprudence. We cannot live the AI evolution in a passive way and is a matter of time before it is adopted in the criminal justice system. The following paper focuses on the respect for fundamental rights as a filter of such instruments. We highlight the right to a fair trial (article 6), the principle of legality (article 7) and the prohibition of discrimination (article 14). Predictability can justify the adoption of predictive tools, ensuring fairer decision. On the other hand, explainability is an essential requirement that has been developed by explainable artificial intelligence. There are several AI models that must be adopted depending on domain and intended purpose. Only a multidisciplinary approach can ensure the compatibility of such instruments with ECHR. Thus, a confrontation between legal and engineering concepts is essential so that we can design tools that are more efficient, fairer and trustable.

Keywords: Artificial intelligence · Criminal system · European Convention on Human Rights

1 The Use of AI Systems as an Aid in Determining Sentence Length

We can divide two main branches of AI (i) symbolic AI and (ii) sub-symbolic AI. The first expresses knowledge by representing it according to rules. In these, the inference mechanisms are solidified and there are several representation languages (as logic or imperative languages). Nevertheless, sometimes there are areas in which we do not have the full extent of knowledge, being necessary to learn based on cases.

Symbolic AI is based on classic logic. In sub-symbolic AI we are faced with the ability to learn based on data. The legal universe is more associated with the symbolic world. Thus, we can think of the adoption of Expert Systems that are based on rules previously defined by experts. However, there are cases in which it is necessary to attend to data. In these scenarios, there is an extraction of knowledge through generalization and there is always an error associated with the model itself. The latter models can help detect inconsistencies between cases, for example.

We cannot choose which branch of AI we will use in the justice domain. In fact, we believe that both will be used, depending on the specific function assigned to such a tool.

G. Marreiros et al. (Eds.): EPIA 2022, LNAI 13566, pp. 108–118, 2022.
https://doi.org/10.1007/978-3-031-16474-3_10

Nevertheless, for aid in determining the measure of the punishment we believe that the analysis of previous decision will be a fundamental step.

As set out above, AI can be applied in the justice systems in various ways. At the investigation level, it can assist in analyzing evidence. It can also help by benchmark legislation. So, there are several possible applications. To understand them, it is necessary to analyze the function for which it can help the judge and how it will change the judge's work.

To fully understand the limitations of AI and the domains in which it can act we appeal to the distinction of (Searle 2002). We can distinguish AI and human performance by resorting to the concepts of syntax and semantics. Computer have syntax, i.e., a formal structure of operation. However, they do not possess semantic, i.e., they do not understand the meaning of these operations. There are no algorithms capable of replacing the judge, actin in a human way. So we do not admit the existence of an AI Judge. However, we consider that there are court function that can be performed by AI, namely analytical functions (such as the distribution of cases, for examples); meanwhile others imply a human dimension that will be very difficult to be compatible with AI (for example, the analysis of guilt).

1.1 Supporting Decision in the Criminal Sentencing

Technology is changing the way of thinking about the criminal law, but also the criminal process. In fact, technology has already changed some aspects of the justice system.

At a first level, technology can be a support tool for the judicial systems. This technology already exists in European systems, namely for enable management of court proceeding allowing, for example, the monitoring of cases online, the delivery of documents or case distribution.

At a last level, we have disjunctive technologies that change the judicial process and the judge's role. In this domain we can think of Artificial Legal Intelligence (ALI), i.e., AI systems capable of providing expert legal assistance or taking decisions (Sourdin 2018, p. 1122). Although predictive tools are not yet used in criminal systems of European countries, it seems that it is a matter of time to their adoption. Currently, the University of Cambridge is testing the Harm Assessment Risk Tool to be used in this domain.[1] Moreover, (Aletras et al. 2016) designed an AI model that can predict ECtHR decision with 79 per cent accuracy.

In fact, some studies reveal that AI systems will become more relevant (Sourdin 2018), although it is not clear the concrete domain where they will be used. We exclude the admissibility of such tools as substitutes for the judge, admitting them as auxiliaries to the jurisdiction task, particularly in the area of determining the length of the sentence in order to combat inconsistency between penalties. Thus, we restrict the scope of our study to the use of AI as a tool at the service of human being, as an aid to the judge. This is a consensual point in the doctrine and guidelines of Council of Europe (CoE).

The use of AI systems as tools to assist the judges in measuring sentencing has been seen as a potential way to ensure efficiency and equality in decisions. The use of such

[1] University of Cambridge Homepage, https://www.cam.ac.uk/research/features/helping-police-make-custody-decisions-using-artificial-intelligence.

tools constitutes a new form of knowledge available to the judge and will change the judicial systems (see. Re, Solow-Niederman 2019).

Most decision forecasting systems involve statistical techniques (Hall et al. 2005, p. 16). It would be interesting to develop systems that also allow translating legal reasoning. This will involve continuous and careful work between legal experts and engineers.[2]

2 The Use of AI in Light of ECHR

Recognizing the emergence of technological mechanisms in the justice system, the CoE and the European Union have developed studies in order to understand this topic.

In its study "Algorithms and Human Rights", the CoE expressed concerns in the field of criminal justice – namely, regarding the fair trial and the due process -, which were answered by the "European Ethical Charter on the use of Intelligence in Judicial Systems and their Environment" (2018).

The European Ethical Charter recognizes the need to encourage the use of instruments that promote the efficiency and quality of justice. Furthermore, the need for such instruments to respect fundamental rights, namely the ECHR, is reinforced.[3] However, the compatibility with human rights will depend on the domain in which it is used and the purpose of it. The choice of system used will be influenced by its functionality, complexity and accuracy (Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems).

Note that the use of algorithms in the judicial system is recognized as a high risk to human rights (paragraph 11 Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems). Thus, the adoption of such instruments should be preceded by a careful study of their risks and benefits. With this in mind the CoE intends the formulation of common framework of standards for the use of AI by courts (cf. Recommendation 2102 (2017). Technological convergence, artificial intelligence and human rights, paragraph 9.2.) and CEPEJ Working Group proposed the creation of a publicity accessible Resource Centre where all AI applications in the field of justice would be registered (cf. CEPEJ (2021)16- Revised roadmap for ensuring an appropriate follow-up of the CEPEJ Ethical Charter on the use of artificial intelligence in judicial systems and their environment).

Following the European studies, we propose to analyze the compatibility of the adoption of AI instruments to help do determine the punishment of the penalty with the ECHR and the ECtHR jurisprudence.

[2] See, *v.g.*, the project of (Hall et al. 2005) about a model of supporting discretionary sentencing decision-making that used Knowledge Discovery from Databases (KDD) to model the discretionary task of the judge.

[3] It should be noted that the present reflection aims at a general analysis. For a particular analysis, we must take into account the specificities of each country and its legislation. As an example, in the Portuguese case, we would have to consider the Portuguese judicial organization and its specific legislation, namely the Portuguese Charter on Human Rights in the Digital Era.

2.1 Right to a Fair Trial

Fairness is the fundamental principle of article 6. This principle requires particular attention in the context of criminal law, which contemplates stricter requirements (Moreira Ferreira v. Portugal (no. 2) [GC], §67; Carmel Saliba v. Malta, §67).

The principle is applicable since the pre-trial stage of the proceedings (inquiry, investigations), because the criminal process is seen as a whole (Dvorski v. Croatia, §76) and the fairness can be harmed since the beginning of the proceedings (Imbrioscia v. Switzerland, §36). It covers the whole proceeding, including the determination of the sentence (Aleksandr Dementyev v. Russia, §23).

The article establishes the right to be heard by an independent and impartial tribunal. Thus, impartiality and independence are the key words. Although predictive tools are not considered “judges” under article 6, it is important to analyze to what extent they may compromise the judicial system.

Independence is evaluated according to different criteria, namely, how judges are appointed, the duration of their term in office, the guarantees against external pressures and the appearance of independence (Fidlay v. United Kingdom, §73). Special emphasis should be given to the requirement that the judge must not be influenced by external pressures (Guðmundur Andri Ástráðsson v. Iceland [GC], § 234), regardless those influences are within the judicial system or outside of it. Specifically, the judge may not receive directives from other judges (Parlov-Tkalčić v. Croatia, § 8). But what if those directives come from AI systems?

Impartiality claims the absence of prejudice or subjectiveness (Kyprianou v. Cyprus [GC], § 118; Micallef v. Malta [GC], § 93). The judge must not attend to his or her personal interests or convictions. Subjectiveness is difficult to prove[4], but it has been pointed out as one of the reasons for the adoption of AI instruments. The use of AI has the potential to be seen as more neutral and reliable than human decisions, as long as they are not opaque (Simmons 2018, p. 1090). But caution is advised, because AI tools can also contain biases.

Given the link between independence and impartiality, we will analyze the two requirements together, similarly to ECtHR (Findlay v. the United Kingdom, §73).

AI can influence the decision of the judge to a point that he or she is strongly inclined to follow its suggestion (Quattrocolo 2020, p. 211). However, if such instruments are based only on previous decisions, in practice the judge is following the decisions of other judges, thus becoming subject to their peers’ influence and biases.

If such tool is to be used, we believe that it should not be based solely on previous cases. The factors that contribute to the decision and the relationship between such factors should be identified, because they help to better understand how the sentence measure is determined, guide the criteria that judges should attend to and ensure sentence consistency across similar cases. Such tools should even detect flaws in previous decisions and correct them. Thus, the judge would not be influenced by peers, but rather rationalize his or her decision, making it more just.

[4] In fact, the ECtHR considers that the training and experience of the judges means that they are not influenced by external pressures (Craxi v. Italy (no.1), §104). For this reason, in most cases the objective aspect is evaluated. Thus, it is evaluated whether the judge has offered guarantees that exclude legitimate doubt as to his or her impartiality (Kyprianou v. Cyprus [GC], § 119).

The court must not only be independent, bust must also appear to be independent. Appearance of independence is important, in that it enables trust in the courts (Şahiner v. Turkey, § 44). For this reason, it must be guaranteed that the final decision maker is the judge, as an authority that can follow, ignore or change the recommendation made by a predictive algorithm (Simmons 2018, p. 1096). However, the judge must justify his or her choice.

The judge must know the system and its limitations. Several studies have pointed out explainability as the most critical. The level of explainability will depend on the application domain[5] and can be required from the moment of creation (by designing systems that are easy to understand, such as decision trees) or with the use of post-hoc techniques (Hamon et al. 2021, p.4). However, it should be noted that explainability is limited by the current state of the art and it should not be demanded of it what is not demanded to human deciders.

Thus, explainability must be taken into account when choosing the AI model used in the judicial systems, because different models have different approaches to this requirement. We can choose models that are considered inherently to be transparent (*v.g.*, linear regression or Bayesian models) or opaque models (*v.g.*, random forests or multi-laser Neutral Networks). Moreover, there are already authors who defend mixing the two models (hybrid models) in order to build explainable and accurate models.[6]

Transparency is different from explainability. A model is transparent if it is understandable by itself (Arrieta et al. 2020, p. 85). Transparency can be analyzed in three dimensions: simulability ("model's ability to be simulated by human"), decomposability ("ability to break down a model into parts and then explain these parts") and algorithmic transparency ("ability to understand the procedure") – cfr. (Belle et al. 2021, p. 3).

The first models are considered to be transparent and can have one or all of the levels of model transparency described. Although considered to be transparent, can become complex and require explainable artificial intelligence (XAI)[7] approaches to explain model decisions (Belle et al. 2021, p. 8).

Opaque models may achieve more accurate results, but their explainability requires the use of XAI, designing explainable models (Heaven 2020)[8], resorting to post-hoc

[5] For example, the use of AI by Spotify does not need an explanation. But if AI is used in areas that can influence human rights we should demand explainability.

[6] For example, (Wan et al. 2021) studied Neural-Backed decision tress (NBDTs) that combines neural networks and decision trees.

[7] There is no consensual definition of XAI. It aims to enable explainability of AI systems, ensuring greater confidence in their use. To understand the concept and the different purposes' of XAI see. (Arrieta et al. 2020).

[8] See the "Explainable AI- Rationale Generation" project which aims to develop machine learning models that automatically generate the machine's inherent reasoning in natural language. In this project, computer scientists have made efforts to justify automated systems, namely, through explanations made in the way that would be done by a human – see. https://gvu.gatech.edu/research/projects/explainable-ai-rationale-generation.

techniques (Belle et al. 2021).[9] By explainability we mean the ability for humans to understand the decision of AI systems.

Thus, it's crucial to choose the right model to develop a decision-support tool. This choice constitutes a relevant moment that will stipulate the level of explainability.

Furthermore, explainability is as important as the way is communicated to the target audience. So, explainability would only be satisfactory if its target audience understands it, which will increase confidence in the use of AI instruments.

The use of AI is not exclusive to its creators or to a fringe of society. AI has expanded to various domains and it is important that its users understand it. It would be interesting to develop studies that specifically target the legal application of such tools, in order to understand the explanations required and how should be legally regulated. AI should not be seen as a bubble, but as a tool to be integrated into various domains that deserves specific thought from each area in which it is used.

Although solutions can be found from XAI, other possible solutions include explaining AI-supported decision making as an alternative or addition to XAI (Bruijn et al. 2022, p. 5). In addition to the explanation of AI tools, we should also require an explanation of the decision that is based, even partially, on AI. In fact, the transparency of the model its different from the transparency of the decision. This implies that judges have the obligation to explain their decision, which forces them to critically analyze the result of AI tools. This requirement guarantees the judge's autonomy and addresses the CoE concern about the effects on the cognitive autonomy of individual (Decl(13/02/2019) - Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, paragraph 9).

Moreover, the accused has the right to an adversarial trial, which means that he has the right to participate effectively in the process by challenging the evidence presented. (Murtazaliyeva v. Russia [GC], § 91). It seems to us that this right allows the parties to challenge all the evidence presented, but also to syndicate the judge's decision and all the factors that contributed to it. Courts decisions must be justified, so that the defendant understands the decision (Moreira Ferreira v. Portugal (no. 2) [GC]. §84) and can exercise the right of appeal.

Additionally, the design of the AI is also important. A mere statistical model does not substantiate a sentence; legal meaning needs to be introduced alongside the empirical data (Reiling 2020, p.8).

2.2 No Punishment Without Law

Article 7 establishes the principle of legality - *Nullum crimen, nulla poena sine lege.* (There is no crime without law. there is no penalty without law). The law must provide in advance the conduct that constitutes a crime and the penalty cannot exceed the limits set (Del Río Prada v. Spain [GC], §80). This principle is important in the stages of prosecution, conviction and punishment (Del Río Prada v. Spain [GC], §77).

[9] There are several types of post-hoc explanations that will be appropriate depending on the model used that should be combined to obtain a more comprehensive explanation (Belle et al. 2021). See also (Arrieta et al. 2020).

The crime and the penalty must be clearly defined in law. According to the interpretation of the ECtHR the term "law" covers not only legislation, but also case law, comprising qualitative requirements, namely accessibility and foreseeability (Cantoni v. France, §29; Del Río Prada v. Spain, §91). Thus, the principle of legality covers both the law and the way the law is applied in a given case, which shows that foreseeability may be a factor to be taken into account.

These requirements apply to both the definition of the offense (Jorgic v. Germany, §§ 103–114) and the penalty. It is important to understand what the courts consider "foreseeability" to understand whether the concept coincides (at least in part) with the predictive justice we are writing about.

The term "predictability" is generally used as a way to ensure legal certainty. This requirement is seen as a counter power to *ius puniendi*. Knowing the possibility of a criminal consequence is different from predicting the concrete case (Quatrocollo 2020, p.219). When analyzing the jurisprudence of ECtHR we do not find express reference to the need to foresee the concrete case. Therefore, the concept does not seem to coincide with that of predictive justice.

When someone has committed a criminal act, it is important that it and its consequences are foreseen in previous law. When committing the criminal act, the agent acts even though he/she knows that the conduct is outside his/her field of freedom, to which a penalty corresponds (Amado 2018). But the agent is not expected to reflect on the penalty concretely applied. Even so, the agent knows that more serious conducts will have more severe penalties, there being a minimum of foresight regarding the scale of his/her penalty within the legal framework provided.

Predictability is different from prediction. Furthermore, it seems to us that the principle of legality requires something more than the mere provision of a criminal consequence in a previous law.

When we understand "law" in such a broad sense, it is not only the legislation that indicates criminal conduct. The ECtHR notes that we should be guided by the courts' interpretation. In the Camilleri v. Malta decision (see. §§39–45) the court refers the predictability of sentencing standards. It seems to us that, with the necessary adaptions, we can consider that, if there are serious and unjustified situations of inconsistency in sentencing, we are faced with a violation of this principle, because it does not correspond to the standard applicable by case law.[10]

This may have a contradictory effect. Judges may begin to adopt more severe penalties in cases that deserve more favorable sentences (Amado 2018, p. 185). We do not agree with this argument insofar as the judge will always have the possibility to decide differently from the applicable standard, if that is objectively justified.

Predictive justice can help predict the appropriate penalty for the case. The design of these tools will help to understand the factors that influence sentencing.[11] For this reason, it seems to us that such a tool could ensure a better application of this precept, by allowing the identification of cases that fall outside the predictable patterns of application in order to subsequently analyze whether this is justified in light of objective consideration.

[10] However, if the consequence is more favorable than the one corresponding to the foreseeable standard, there is no violation of the principle (Amado 2018, p. 180).

[11] In the opposite direction (Quattrocolo 2020, p. 219).

When it comes to penalties, the courts, faced with the penal framework, define the actual penalty for a case according to legal criteria. So, similar cases should have similar penalties. But if the penal frames are too broad, we could have a greater openness and discretion on the part of the judge, which could lead to unpredictability of penalties.

The ECtHR has a subsidiary nature in this matter and because of that cannot analyze the error of fact or law, unless the national court has violated rights and freedoms of ECHR (Vasiliauskas v. Lithuania [GC], §189).

The court cannot interfere in matters of determining the measure of the sentence. According to article 7, the court must confine to see whether the penalty imposed is not more severe than the penalty provided at the time of the practice of the fact and if the principle of retrospective application of more favorable criminal law was respected. However, the ECtHR cannot assess the length or type of applicable penalty (Vinter and Others v. the United Kingdom [GC], §105). However, we must not forget that the ECtHR can assess compliance with the ECHR. Thus, if there is a manifestly disproportionate penalty, this can be considered by the court under article 3 (Vinter and Others v. the United Kingdom [GC], §102) and, in our opinion, penalties that deviate from the standard application of the courts may also violate article 7.

2.3 Prohibition of Discrimination

The right not to be discriminated against complements the other articles of the ECHR and the Protocols. The article 14 is complemented by article 1 of Protocol No. 12 which establishes a general prohibition of discrimination. Both articles prohibit direct and indirect discrimination.

The application of this principle refers to the rights and freedoms of the ECHR, so its violation must be analyzed with another provision (Inze v. Austria, §36), which is why we refer to the rights analyzed above.

If AI tools are used in the criminal system, the right to a fair trial (article 6) must be guaranteed. This right will be violated if the decision takes discriminatory factors into account. The algorithm will be based on a theory (in the form of determining the penalty) translated into a code and also in data (depending on whether one opts for data-driven regulation or code-driven regulation, although it seems to us that both is ideal).

Although it is argued that the use of AI systems will allow the fight against human subjectivity, in fact, that may reveal the subjectivity of their creator or even discriminatory aspects reflected in their data. Therefore, we must be careful when building the algorithm and defining the data to be used. Some forms of AI that are currently used have proved this risk (Sourdin 2018, p. 1129). Recently, several studies have indicated the discriminatory aspects and biases that the COMPAS (an AI system used in US courts) suffers, which calls into question the very usefulness of the tool (Freeman 2016). When analyzing the data, if a rigorous choice is made, flaws may be detected in the justice system that would not be detected otherwise, but if data is not carefully chosen, it can result in discriminatory decisions that are intended to be countered.

Note that direct and indirect discrimination are prohibited. If the seemingly neutral AI discriminates based on the relationship with a group of people, there is indirect discrimination. So, it is important to ensure the quality and integrity of the used data.

However, the discrimination may not exist from the start, but results from machine learning. So, it is important to analyze the risk of bias throughout its use.

Therefore, its fundamental to understand the algorithm to detect these issues. In this respect, XAI could help to highlight bias in data (Arrieta et al. 2020)[12] and reverse engineering or reverse control is referred to as a solution that allows to review, discuss and contest the results (Quattrocolo 2019, p. 1527).

Furthermore, once ensuring respect for fundamental rights is an ongoing task, impact assessments must be carried out before and during the use of such tools.

3 Conclusions

The design and the implementation of AI systems must be compatible with fundamental rights and any discrimination must be prevented. Certified sources must be used; transparency, impartiality and fairness must be guaranteed. This will depend, to a large extent, on the algorithmic regulation and data processing model, which is why the tools must be designed in a multidisciplinary way.

Although XAI is quite recent, we believe that can make relevant contributions regarding explainability, which is the key concept here. It is important to understand the advances of AI in order to analyze its future integration into criminal justice. AI should be evaluated not only for its usefulness but also for the process it uses to achieve them.

The XAI research points out that AI systems should be used in more domains and their users should be part of the design from the very beginning because different people need different kind of explanations. In fact, we should think in explainability since the design: the required explanations must be defined and the model should be designed to provide the desired results and comply with the requirements demanded by law.

Throughout this paper we have noted that are several questions which have to be answered in collaboration with AI specialists and legal experts. An approach that takes into account the criminal law specificities and attempt fundamental rights is required. This knowledge exchange will create a link between XAI and legal world.

Engineering and social sciences should join efforts to establish metrics regarding the level of explainability required since this will be the guarantor of fundamental rights. This multidisciplinary approach requires continuous monitoring, since today's issues may be outdated tomorrow. Indeed, application of AI may be denied today, but as the capabilities of AI improves and confidence in it increases, it may be accepted.

It seems to us that only after testing several models can we state which one should be adopted and, consequently, which XAI solutions may be necessary. Thus, many questions will remain unanswered and should be taken up again when analysing and testing the concrete models. However, the adoption of these instruments must comply with the ECHR. We should anticipate interpretation/requirements made by ECtHR, given the particular aspects arising from the use of AI instruments in criminal proceedings.

Finally, engineering concepts do not coincide with legal concepts. For example, the efficiency sought by justice, in this particular case, the consistency of penalties, does

[12] For example, XAI techniques can be used to identify hidden correlations between data – see. (Arrieta et al. 2020, p. 104 ff.).

not have the same meaning as the efficiency sought by AI (Quattrocolo 2019, p. 1531). Thus, in the design of these models, a multidisciplinary conception that translates legal reasoning and guarantees fundamental rights is necessary. In fact, we must design reliable AI instruments that focus on human rights.

Several studies focus on the technical evolution of AI tools, making the use of such tools dependent on the evolution of computer science. For example, the demand for explainability is the fundamental issue in any discourse on the application of AI systems in the criminal justice system. This requirement depends on the evolution of AI systems and their transparency. However, this is an issue that does not exclusively concern engineering. It is a multidisciplinary field that should have at its core the respect for fundamental rights as a filter of these tools.

Acknowledgement. This research was funded by the Portuguese Foundation for Science and Technology (FCT) through the PhD grant 2021.07986.BD.

FCT Fundação para a Ciência e a Tecnologia REPÚBLICA PORTUGUESA | CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR UNIÃO EUROPEIA Fundo Social Europeu

References

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. Peer J. Comput. Sci. **2**, e93 (2016). https://doi.org/10.7717/peerj-cs.93

Amado, J.A.G.: On the principle of criminal legality and its scope: foreseeability as a component of legality. In: Pérez Manzano, M., Lascuraín Sánchez, J.A., Mínguez Rosique, M. (eds.) Multilevel Protection of the Principle of Legality in Criminal Law, pp. 177–193. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-63865-2_10

Arrieta, A., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsabile AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012

Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. Front. Big Data **4**, 688969 (2021). https://doi.org/10.3389/fdata.2021.688969

Bruijnm, H., Warnier, M., Janssen, M.: The perils and pitfalls of explainable AI: strategies for explaining algorithmic decision-making. Gov. Inf. q. **39**(2), 101666 (2022). https://doi.org/10.1016/j.giq.2021.101666

Freeman, K.: Algorithmic injustice: how the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. North Carol. J. Law Technol. **18**(3), 75 (2016)

Georgia Tech Homepage. https://gvu.gatech.edu/research/projects/explainable-ai-rationale-generation. Accessed 28 Apr 2022

Hall, M., Calabrò, D., Sourdin, T., Stranieri, A., Zeleznikow, J.: Supporting discretionary decision-making with information technology: a case study in the criminal sentencing jurisdiction. Univ. Ottawa Law Technol. J. **2**(1), 1–36 (2005)

Heaven, W.: Why asking an AI to explain itself can make things worse. MIT Technology Review (2020). https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse/. Accessed 28 Apr 2022

Hamon, R., Junklewitz, H., Malgieri, G., Hert, P., Beslay, L., Sanchez, I.: Impossible explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. In: Proceeding of ACM FaaCT. ACM, New York (2021). https://doi.org/10.1145/1234567890

Quattrocolo, S.: Artificial Intelligence, Computational Modelling and Criminal Proceedings. A Framework for a European Legal Discussion. Springer, Heidelberg (2020)

Quattrocolo, S.: An introduction to AI and criminal justice in Europe. Revista Brasileira de Direito Processual Penal **5**(3), 1519–1554 (2019). https://doi.org/10.22197/rbdpp.v5i3.290

Reiling, A.: Courts and artificial intelligence. Int. J. Court Adm. **11**(2), 1 (2020). https://doi.org/10.36745/ijca.343

Re, R., Solow-Niederman, A.: Developing artificially intelligent justice. Stanford Technol. Law Rev. **22**(2), 242–289 (2019)

Searle, J.: Can computers think?. In: Chalmers, D.J. (ed.) Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press (2002)

Simmons, R.: Big data, machine judges, and the legitimacy of criminal justice system. UC Davis L. Rev. **52**, 1067–1118 (2018)

Sourdin, T.: Judge v Robot? Artificial intelligence and judicial decision-making. UNSW Law J. **41**(4), 1114–1133 (2018)

Wan, A., et al.: NBDT: neutral-backed decision tree. In: ICLR Conference Paper (2021). https://doi.org/10.48550/arXiv.2004.00221