Taking Defence Rights Seriously. The Need for Explainable Artificial Intelligence in Criminal Proceedings – A European Perspective

Marcia Michalina
University of Wrocław
ORCID 0000-0002-7872-8507

DOI: 10.54103/milanoup.215.c463

Abstract

The integration of artificial intelligence (AI) into criminal proceedings introduces both opportunities and challenges, also regarding the protection of defence rights. This article examines the intersection of AI, the right of defence and procedural fairness from a European perspective. It presents the hypothesis that the current deployment of AI in criminal justice systems presents more risks than benefits to defendants, particularly due to issues such as the black box problem and limited explainability of AI systems. The research explores how inadequate practice of AI systems implementation undermines key rights, including the right to information, defence preparation, legal assistance, translation and also equality of arms. The objectives include assessing whether the AI application can infringe defence rights and in what way, identifying the need for explainable AI (xAI), and proposing solutions to mitigate the risks posed by AI use in criminal trials. The methodology involves a review of European legal frameworks (regarding both defence rights standards and AI), case law, and practical examples of AI applications in criminal proceedings, supported by analysis of technical and human rights implications of AI systems like predictive analytics, natural language processing, and image recognition. Findings highlight critical gaps in regulation, emphasising the necessity for xAI to ensure transparency, effective participation, and trust in judicial processes. Recommendations include adopting a further harmonised legal framework for AI in criminal proceedings and implementing explainability measures tailored to defendants' needs.

Keywords
Artificial Intelligence, Right of Defence, xAI, Criminal Proceedings

1. Introduction

The technologically-based measures have deeply affected criminal proceedings in recent years. The need for digitalization was demonstrated by the COVID-19 spread and dissemination of remote court hearings.[1] Many controversies were also raised by surveillance technologies such as Pegasus, online searches and new lie-detecting technologies.[2] One of the most advanced technologies that is believed to significantly modify the course of proceedings is Artificial Intelligence (AI). Although the AI use has been mainly connected with and developed in the private law,[3] recently the topic has gained significant interest from the public law perspective. The use of novel technology in the criminal proceedings, including some of the AI systems, appears to be inevitable, especially due to possible advantages of their implementation.

Application of AI systems comes with both risks and benefits. AI-based systems are believed to facilitate many activities taking place during the course of proceedings. They can make them more effective, efficient, quicker and eliminate bias rooted in the human factor.[4] The most commonly describes risks are connected with bias and right to privacy, but they can also negatively influence

1 Krešimir Kamber, 'The Right to a Fair Online Hearing' (2022) 22 (1) Human Rights Law Review 1, 'The Practice Magazine' (*Harvard Law School Center on the Legal Profession*) <https://clp.law.harvard.edu/knowledge-hub/magazine/> accessed 20 March 2023, Ilze Tralmaka 'Defence Rights in Remote Justice Procedures' (*UN Office on Drugs and Crime)* <https://www.unodc.org/dohadeclaration/en/news/2020/06/defence-rights-in-remote-justice-procedures.html> accessed 30 May 2024.

2 Orin S Kerr, 'Digital Evidence and the New Criminal Procedure' (2005) 105 Columbia Law Review 279, Michele Simonato, 'Defence rights and the use of information technology in criminal procedure' (2014) 85 Revue internationale de droit pénal 261, Sunil Bector, '"Your Laptop, Please:" The Search and Seizure of Electronic Devices at the United States Border' (2009) 24 Berkeley Technology Law Journal 695, Bruce Luber and others, 'Non-Invasive Brain Stimulation in the Detection of Deception: Scientific Challenges and Ethical Consequences' (2009) 27 Behavioral Sciences & the Law 191.

3 Commission, Directorate-General for Justice and Consumers, *Liability for Artificial Intelligence and Other Emerging Digital Technologies. Report from the Expert Group on Liability and New Technologies – New Technologies Formation* (Publications Office of the European Union 2019), Jochen Hanisch, *Haftung für Automation* (Cuvillier 2010), Michael Faure, Louis Visscher and Franziska Weber, 'Liability for Unknown Risks – A Law and Economics Perspective' (2016) 7 Journal of European Tort Law 198, Eric Hilgendor and Susanne Beck (eds) *Robotik und Gesetzgebung. Beiträge der Tagung vom 7. bis 9. Mai 2012 in Bielefeld* (Nomos 2013).

4 Sray Agarwal and Shashin Mishra, 'Introduction' in Sray Agarwal and Shashin Mishra (eds), *Responsible AI: Implementing Ethical and Unbiased Algorithms* (Springer International Publishing 2021), Daniel Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence* (Basic Books 1993).

other human rights, including rights of defence, on which the article will focus.[5] The main hypothesis of this article is that currently the use of systems based on the AI in criminal proceedings is more of a threat than a benefit to the right of defence.

Due to the use of AI systems and lack of sufficient regulation on explainable intelligence, right of defence could be significantly restricted, especially with regard to the right to information and the black box problem.[6] Right to information, to prepare the defence and to an active participation in the proceedings is not effective when gathered evidence materials are incomprehensible and the decision-making process is unclear. The current state of regulation on the European Union (EU) level and in majority of Member States still does not sufficiently safeguard the right of defence and does not provide adequate means for review.[7] This can lead to aggravating data bias and limiting the quality of individual rights. The use of defence-assisting systems is also underregulated. Although the black box problem has been addressed several times, some examples of explainable AI (xAI) have appeared,[8] and even some attempts to regulate the issue have been made,[9] the article argues that there are still many practical difficulties with an effective execution of procedural rights. Therefore, the need for regulation and safeguarding explainability, especially considering criminal proceedings specifics, is still valid.

The article focuses on the European perspective. More specific approach to the problem can lead to presenting more precise conclusions. It will allow to analyse the scope of defence rights applicable to the particular system and establishing the point of reference. Due to many differences among the particular regions in this regard (eg between US, EU, regional and universal ones), clarifying the chosen standard is crucial as it may result in recreating the coherent set of rights and adapting them to the new technological demands.

The first part of the article focuses on defining the AI and determining general applications and problems caused by the AI use in the defence rights context. The existing definitions on the AI and basic elements of the AI systems, such as machine learning and neural networks are briefly analysed. The article further introduces the concept of the black box problem and xAI – methods

5 Michele Caianiello, 'Criminal Process Faced with the Challenges of Scientific and Technological Development' (2019) 27 European Journal of Crime, Criminal Law and Criminal Justice 267.

6 For the definition see below – s 2.3.

7 Center for AI and Digital Policy, 'Artificial Intelligence and democratic values index' <https://www.caidp.org/reports/aidv-2021/> accessed 30 May 2024.

8 Stephen K Reed, 'Explainable AI' in Stephen K Reed (ed), *Cognitive Skills You Need for the 21st Century* (OUP 2020), Joanna J Bryson, 'The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation' in Markus D Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (OUP 2020).

9 eg in the AI Act and the Proposal for the AI Act (see further para 2.3).

of its achievement, scope and attempts to regulate the matter – as well as the practical risks connected with the black box, especially due to the proven track of bias of the AI systems. In this part the research is mainly based on the existing literature and analysis of legal acts, especially the Artificial Intelligence Act[10], supported by the empirical research elements with regard to publicly available information on the systems used.

The next part covers the examples of AI systems and possibilities for their application in the criminal proceedings. It presents the general types of AI systems that are or can be used in criminal proceedings. Then the contribution explores, how they can affect the right of defence. The main basis for recreating the EU standard on the right of defence is the European Convention on Human Rights (ECHR)[11] and case law of the European Court of Human Rights (ECtHR). It creates a minimum European standard for all Council of Europe members, setting an optimum reference point for analysis also from the EU perspective.[12]

Finally, the article aims to propose possible solutions to the described problems, answer the main research question, if the current state of regulation safeguards the defence rights to the sufficient extent. Its objective is also to analyse if explainability is needed to safeguard effectiveness of defence rights and which ways to achieve xAI can be most beneficial to the defendant. Then the article seeks to answer if issuing legally binding act on the EU level is needed and what guarantees or requirements should be introduced by the legislator.

2. Artificial Intelligence and xAI

2.1 Defining AI and AI systems

Establishing one, common definition of the artificial intelligence has been one of the most problematic issues of the AI-focused literature and practice. Currently there exist many definitions of this term that very much differ. One

10 European Parliament, Legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), P9_TA(2024)0138, hereinafter as the 'AI Act'.

11 Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

12 The 'Explanations relating to the Charter of Fundamental Rights of the European Union' [2007] (OJ C 326) on the arts 47 and 48 (right to an effective remedy and to a fair trial, presumption of innocence and right of defence) directly refer to the ECHR. Also, according to the art 6(3) of the Consolidated Version of the Treaty on European Union [2016] OJ C202/1, fundamental rights, as guaranteed by the ECHR and as they result from the constitutional traditions common to the Member States are the element of the EU law principles.

of the major works on collecting the existing definitions of AI was made by the Joint Research Centre (JRC) of the European Commission which proposed an operational definition of AI formed by 'a concise taxonomy and a set of keywords that characterize the core and transversal domains of AI'.[13] These common characteristics would be consideration of the real world complexity, information processing, decision making, achievement of specific goals.[14] Another definition in EU was proposed by the High Level Expert Group on Artificial Intelligence, which was much more complex and described AI systems as 'software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions'.[15]

The existing definitions still raise many doubts as to their scope and content.[16] There even have appeared statements that creating and adopting one common definition of AI is not needed at all and can be seen as restricting and unnecessary.[17] The definitions that are being constructed in the upcoming regulations tend to approximate and focus on some common elements that constitute the AI in the general comprehension. The AI Act defines AI system as a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs

13 Sofia Samoili and others, *AI Watch: Defining Artificial Intelligence: Towards an Operational Definition and Taxonomy of Artificial Intelligence* (Publications Office of the European Union 2020); Sofia Samoili and others, *AI Watch, Defining Artificial Intelligence 2.0: Towards an Operational Definition and Taxonomy for the AI Landscape* (Publications Office of the European Union 2021).

14 Gianluca Misuraca and Colin Van Noordt, *AI Watch - Artificial Intelligence in Public Services* (Publications Office of the European Union 2020).

15 Independent High Level Expert Group on AI (AI HLEG), 'A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines | Shaping Europe's Digital Future' (8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>. It also specified that as a scientific discipline, AI includes 'several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)'.

16 Joanna J Bryson, 'Europe Is in Danger of Using the Wrong Definition of AI' *Wired* <https://www.wired.com/story/artificial-intelligence-regulation-european-union/> accessed 4 April 2023.

17 Manuel A Utset, 'Predictive Policing and Criminal Law', in John McDaniel and Ken Pease (eds), *Predictive Policing and Artificial Intelligence* (Routledge 2021).

such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law[18] adopts a very similar definition and describes AI system as a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments. The Conventions states that different AI systems vary in their levels of autonomy and adaptiveness after deployment.

It became clear that there are some basic AI-specific elements that differentiate it from other common algorithms and represent its basic core. [19] It is, first of all, the ability to learn, reason and make decisions. The aim of the reasoning, to put it simply, is to present solutions of the defined problems without the human intervention. It is an essential part of the knowledge representation and reasoning part of the AI. The decision–making part is an element of the process that allows to obtain a final result and reach the aim of the AI systems use. The part mostly differentiating AI systems from ordinary algorithm–based ones is an ability to learn. To simplify, the self-learning program should, on the basis of the input data entered into the system and in connection with the expected output, modify and shape the rules for processing data itself, and thus learn through its experience. Several types of machine learning can be distinguished that demand various degrees of human impact.

Supervised learning allows to determine the relationship of specific inputs to outputs based on labelled examples of pairs of both types of data provided to the system. The implemented patterns can then provide examples for future use, based on the similarity of the situation, environment, etc. Supervised learning most often requires a significant amount of data and human action, especially data labelling.[20] Unsupervised learning features a greater degree of autonomy in the AI systems. The input data is not subject to prior labelling, and the system itself aims to find patterns in the datasets. The AI–based system is supposed to find the links that exist between the data, which should be identified with increasing accuracy as the system continues to operate. As a result it should achieve the goals set for the AI use. Systems in this case can function

18 Committee of Ministers of the Council of Europe (133th Session), 'Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law' (17 May 2024 Strasbourg), that will be opened for signature on the occasion of the Conference of Ministers of Justice in Vilnius on 5 September 2024.

19 This approach is the most similar to the definition provided in Sofia Samoili and others, *AI Watch* (n 13) and Sofia Samoili and others, *AI Watch 2.0.* (n 13).

20 Pádraig Cunningham, Matthieu Cord and Sarah Jane Delany, 'Supervised Learning' in Matthieu Cord and Pádraig Cunningham (eds), *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (Springer 2008).

more independently and rely less on human intervention.[21] The semi-supervised learning combines elements of the two types of machine learning mentioned above - both labelled and unlabelled input data are introduced to the system.[22]

There can be also distinguished other specific types of machine learning, for example reinforcement learning, which relies on the system's search for solutions and ways of data processing to maximize the result in a given environment. The system provides for the environment analysis, most often subject to many variables, and by learning through its experience aims to achieve the desired effect.[23] Finally, there is deep learning, that does not necessarily constitute different mode of machine learning, but is strictly related to the type of system and its basis on which machine learning is based in this case - artificial neural networks.

2.2 Artificial neural networks

The basis of many modern artificial intelligence-based systems and deep learning algorithms are formed by neural networks, also referred to as artificial neural networks (ANNs).[24] The term neural network refers to the way the human brain functions.[25] The basic unit in this case is the so-called node, an artificial neuron, which is provided with an input, an output and a dataset storing weights.[26] Each of these nodes connects to the others, collects the input signals, recalculates them and sends the output signal to the next layer. The artificial neuron has a specific associated weight and threshold (connection - hence connectionist AI). If the output data exceeds the assigned threshold value, the specified node will be activated and the data will be sent to the next layer of the network.[27]

---

21 Salim Dridi, 'Unsupervised Learning - A Systematic Literature Review' (13 June 2024) OSF Preprints <https://osf.io/preprints/osf/mpkht> accessed 30 May 2024.

22 Jesper E van Engelen and Holger H Hoos, 'A Survey on Semi-Supervised Learning' (2020) 109 Machine Learning 373, Y Reddy, Viswanath Pulabaigari and Eswara B, 'Semi-Supervised Learning: A Brief Review' (2018) 7 International Journal of Engineering & Technology 81.

23 Wolfgang Ertel, 'Reinforcement Learning' in Wolfgang Ertel (ed), *Introduction to Artificial Intelligence* (Springer International Publishing 2017).

24 Akash Goel, Amit Kumar Goel and Adesh Kumar, 'The Role of Artificial Neural Network and Machine Learning in Utilizing Spatial Information' [2022] Spatial Information Research <https://doi.org/10.1007/s41324-022-00494-x> accessed 30 May 2024.

25 Ralph Adolphs, 'Cognitive Neuroscience of Human Social Behaviour' (2003) 4 Nature Reviews Neuroscience 165.

26 N Jindal and V Kumar, 'Enhanced Face Recognition Algorithm Using PCA with Artificial Neural Networks' (2013) <https://www.semanticscholar.org/paper/Enhanced-Face-Recognition-Algorithm-using-PCA-with-Jindal-Kumar/d7244f2ab95483b68b97a346e40ebf6ad919e1e7> accessed 29 April 2023.

27 'AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?' (*IBM* 6 July 2023) https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks accessed 30 May 2024.

The functioning of neural networks is based on the layers within which the neurons are arranged. The outer layer registers data inputs from outside and processes them. It then passes this information to the next layer, which is already referred to as a hidden layer. The hidden layer recalculates and processes the acquired information and then also passes the information to the next hidden layer or output layer. The output layer processes the aggregated data and information from the previous layers, makes the final calculations and sends the result to the user, which takes the form of making a specific decision.[28]

The use of neural networks allows for higher efficiency of the AI systems and processing huge amounts of information. Combined with machine learning, especially deep learning, constitutes the biggest advantages of the AI and can result in spectacular effects, exceeding human possibilities. However, it also causes some fundamental problems in the AI ethics domain.

2.3 Explainable AI

One of the main issues connected with the use of AI is the black box problem. Its essence comes down to the inability to thoroughly track how an AI-based mechanism produced a given result and tracing its inner processing.[29] The black box is in fact mainly associated with machine learning, especially unsupervised and neural networks based systems. In the case of complex mechanisms based on machine learning, input data and output datasets are fed into the algorithm, which, in combination, are supposed to lead the algorithm to develop optimal methods for achieving its designated goals. This process, however, is not based on explicit, clear rules, adapted to the typical ways and capabilities of human reasoning.[30] The hidden layers of nodes process input data and pass partial output data onto the next layer. The way the systems function does not allow for the results of the processes between the layers to be fully tracked. It is only possible to observe the final output data, which does not allow to learn how the system itself works and how it makes decisions.[31] So, we have a self-learning algorithm that issues recommendations, makes analysis or

28 Richard E Neapolitan and Xia Jiang, *Artificial Intelligence: With an Introduction to Machine Learning, Second Edition* (2nd edn, Chapman and Hall/CRC 2018).

29 Carlos Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34 Philosophy & Technology 265.

30 Gabriëlle Ras, Marcel van Gerven and Pim Haselager, 'Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges' in Hugo Jair Escalante and others (eds), *Explainable and Interpretable Models in Computer Vision and Machine Learning* (Springer International Publishing 2018).

31 Michael Pfeiffer and Thomas Pfeil, 'Deep Learning With Spiking Neurons: Opportunities and Challenges' (2018) 12 Frontiers in Neuroscience <https://www.frontiersin.org/articles/10.3389/fnins.2018.00774> accessed 28 March 2023; Vanessa Buhrmester, David Münch and Michael Arens, 'Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey' (2021) 3 Machine Learning and Knowledge Extraction 966.

predictions that serve as a basis for the final decision of a competent authorities in criminal proceedings that can be unexplainable. As a result, it can cause some serious problems within the right of defence.

There has been a long debate over the issue of explainability and different approaches to this problem have been distinguished. For example, explainability can be supported by providing the broad information about the system used which is not decision or input-data specific (so called global interpretability). This information can include setup information, training metadata, performance metrics, process information.[32] Such explainability can guarantee the proper quality control, especially over the input data and allow defendant to check for eg implemented anti-bias procedures.

The local interpretability (subject – centric as it mainly concentrates on the subject of the decision or recommendation) models can provide information about the characteristics of individuals who received similar decisions.[33] The interpretations are based on the input data and do not give the explanations about the model as a whole, but narrow it down to the chosen variables.[34] For example, the convicted person could receive real information on what elements influenced the penalty imposed.

The decompositional approach on the other hand attempts to explain inner workings or replicate reasoning of the system. One of the most common methods is creating the surrogate model, that does not have access to the inner workings of the primary AI model and internal weights of the model, but works by analysing featured input and output data pairs. A model is constructed based on modelling the response of the simulator to a limited number of intelligently chosen data points.[35] In this case defendant would presumably be able to learn about how the system came to its conclusions.

32 Ashley Deeks, 'The Judicial Demand for Explainable Artificial Intelligence' (2019) 119 Columbia Law Review 1835, Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (23 May 2017) [2017] D*uke* L*aw* & T*echnology* R*eview* 18–84.

33 Edwards and Veale (n 32) 57–59, Danielle Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89 Washington Law Review 1, 28–29.

34 Edwards and Veale (n 32) 59.

35 W Andrew Pruett and Robert L Hester, 'The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes' (2016) 11 PLoS ONE e0156574, Deeks (n 32) 1838.

The explainability issues have been raised in almost every recommendation of international bodies, including Council of Europe[36], OECD[37], and UN[38]. In the most recent years such recommendations have been transferred into regulatory proposals. Transparency obligations have been introduced to the Council of Europe Convention on AI and reflected in the UN General Assembly Resolution[39]. More explicit and specific attempts to introduce some explainability-oriented solutions can be found in the AI Act. The Regulation contains provisions relating to the input data quality control and guidelines on designing and developing the systems, maximizing the traceability. High-risk AI[40] systems should enable the automatic recording of events ('logs') while the systems is operating throughout its lifecycle. Moreover, the AI systems provider should ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. Chapter 3 of the AI Act obliges, inter alia, to draw up technical documentation, conformity assessment and quality management system.

Finally, the high-risk AI systems should guarantee the possibility for human oversight and that they can be effectively overseen by natural persons during the period in which the AI system is in use. Human oversight is one of the postulates frequently appearing in the recommendations addressing the AI ethics domain.[41] This oversight, according to the AI Act, should aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose.

The Proposal for the AI Act[42] introduced obligations that the provider was expected to fulfil for the benefit of the user. It defined 'user' as 'any natural or

36 Council of the European Union, 'Shaping Europe's Digital Future' (Conclusions, 9 June 2020), idem, 'Artificial intelligence: Presidency issues conclusions on ensuring respect for fundamental rights' (Oct. 21, 2020), idem, 'The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change' 11481/20 (Presidency Conclusions, 21 October 2020).

37 Global Privacy Assembly, 'Adopted Resolution on Accountability in the Development and Use of Artificial Intelligence' (October 2020) <https://globalprivacyassembly.org/wp-content/uploads/2020/11/GPA-Resolution-on-Accountability-in-the-Development-and-Use-of-AI-EN.pdf> accessed 30 May 2024.

38 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2021), UNSG 'Roadmap for Digital Cooperation' (June 2020).

39 UNGA, Res 78/L.49 (13 March 2024) UN Doc A/78/L.49.

40 See in particular the Annex III to the Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts', COM (2021) 206 final (hereinafter also as 'Proposal for the AI Act').

41 Council of Europe, Committee of Ministers, *Ad hoc* Committee on Artificial Intelligence (CAHAI), 'Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law' CM(2021)173-add (17 December 2021).

42 See Proposal for the AI Act.

legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity' (art 3(4)). According to this definition, the court or the law enforcement agencies could be perceived as users. However, when the State (the court, prosecutor etc.) would introduce the results of AI systems into the proceedings, there were still no sufficient safeguards for the person affected by the use of such systems – the defendant.

In the final text of the AI Act, due to the amendments proposed by the European Parliament, the position of the defendant in the light of the explainability requirements, shifted and improved to some extent. The AI Act introduces a definition of the 'deployer' as a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity. This allows the term to cover both State and private entities that implement the AI systems directly into their activities. In the amendments of the European Parliament, there also appeared a term 'affected person' defined as 'any natural person or group of persons who are subject to or otherwise affected by an AI system'. The EU legislator clearly saw the need to differentiate and regulate the situation of these two groups of entities.

According to art 13 of the AI Act, high-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. Types and degrees of transparency should correspond with the relevant obligations of the provider and deployer set out in Section 3 of the Act. The article also imposes an obligation to provide the deployers with a detailed information on the AI systems. This allows for making the primary step towards the explainability - it is first of all necessary for the deployers to understand and interpret the AI system to give any real information to the person affected by its use.

The AI Act introduces also the next step – the right to explanation of individual decision-making. According to art 86, as a rule, any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights should have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.

The AI Act introduces, therefore, an enforceable right directly attributable to a person affected by the algorithm, potentially also a defendant. It creates a step towards at least a local explainability. Based on this provision defendants should possess a real ability to gain insight into the decision taken and main parameters of the AI system involvement. This provision, however, can be subjected to exceptions (art 86(2)) and has not necessarily been adjusted to the

criminal proceedings specifics. The final wording of the AI Act also deprived the affected person the explicit insight into the input data. Also, without the national transposition, the awareness of such rights can be in fact minimal. Finally, the provision does not constitute a solution to all problems connected to lack of real explainability of AI systems and information about the reasoning of the system.

3. AI in criminal proceedings

3.1 Examples of the systems used in criminal proceedings

The AI systems, in particular in pre-trial stage, can be used for the purpose of individual and general prevention and policing.[43] Data obtained through the use of AI could be also applied as evidence that would be gathered by the law enforcement agencies and used in the further course of the proceedings.[44] When it comes to issues connected primarily with the judicial stage of the proceedings, the biggest controversy is, of course, the possibility of decision-making by intelligent agents.[45] However, much more likely and common than the vision of judge-robot's independent judgments are systems that assist in reaching a certain type of decision. They can also signal possible omissions, mistakes or propose and calculate, for example, the elements of the punishment assessment based on the data entered by the judge.[46] High hopes for streamlining the course of the criminal proceedings also associated with speech recognition systems and the translation of evidence using AI.[47] AI systems use can be used in different dimensions, including natural language processing (NLP), image recognition and predictive analysis. Although they often intersect, each of them

43 See eg Stefanie Hänold, 'Profiling and Automated Decision-Making: Legal Implications and Shortcomings' in Marcelo Corrales, Mark Fenwick and Nikolaus Forgó (eds), *Robotics, AI and the Future of Law* (Springer Singapore 2018) 123–153.

44 Sabine Gless, 'AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials' (15 May 2020) <https://papers.ssrn.com/abstract=3602038> accessed 28 March 2023.

45 Tony Ho Tran, 'China Created an AI "Prosecutor" That Can Charge People with Crimes' (*Futurism*) <https://futurism.com/the-byte/china-ai-prosecutor-crimes> accessed 28 March 2023.

46 Aleš Završnik, 'Criminal Justice, Artificial Intelligence Systems, and Human Rights' (2020) 20 ERA Forum 567–583.

47 Blessing Oyetunde, 'Introducing Salme, Estonian courts' speech recognition assistant', (*e-Estonia*, 26 January 2022) <https://investinestonia.com/introducing-salme-estonian-courts-speech-recognition-assistant/> accessed 28 March 2023, European Union Agency for Fundamental Rights (FRA), 'Artificial Intelligence, Big Data and Fundamental Rights Country Research Estonia, 2020', <https://fra.europa.eu/sites/default/files/fra_uploads/fra-ai-project-estonia-country-research_en.pdf> accessed 28 March 2023.

can find different application and have their own specifics affecting the needs of the parties to the proceedings.

3.1.1 Natural language processing

Natural language processing (hereinafter: NLP) combines computational linguistics with statistical models, machine learning, including deep learning. It enables computational mechanisms to process human language in the form of text or sound and 'understand' its meaning, including even the intentions and feelings of the author of the text, or the person delivering it.[48] Programmes using NLP perform tasks such as detecting and segmenting speech or text, tagging parts of it, encoding meanings, delivering references, and generating text themselves.[49]

NLP techniques are used in various types of data analysis tools, to detect and dispose of redundant information, to create virtual assistants or chat bots, and to abbreviate, analyse text and perform machine translation. AI systems using NLP are able to process huge amounts of data in a very short time, handle a significant number of grammatical rules specific to different languages, and perform simultaneous translation. NLP programmes are also used for information gathering, search and retrieval, evidence processing and analysis, as well as for question-answering software.[50]

They also offer extensive opportunities in the legal field. It is on natural language processing, especially in the form of Large Language Models (LLMs), that assisting chatbots are based, such as ChatGPT,[51] Automio,[52] BillyBot,[53] tools that search for information and answer legal questions through vast resources of databases, such as Ross,[54] DoNotPay,[55] assisting in challenging fines, addressing small claims or discrimination cases, and the LISA Robot Lawyer,[56] among others. NLP can assist both individuals without legal knowledge and professional lawyers who can use these tools to reduce the time spent on

48 Irene Solaiman and Christy Dennison, 'Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets' (*arXiv*, 23 November 2021) <http://arxiv.org/abs/2106.10328> accessed 20 April 2023.

49 Brojo Kishore Mishra and Raghvendra Kumar (eds), *Natural Language Processing in Artificial Intelligence* (Apple Academic Press 2020).

50 Ivano Lauriola, Alberto Lavelli and Fabio Aiolli, 'An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools' (2022) 470 Neurocomputing 443–456, Pascal Muam Mah, Iwona Skalna and John Muzam, 'Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0' (2022) 12 Applied Sciences 9207.

51 <https://openai.com/blog/chatgpt> accessed 28 March 2023.

52 <https://firmsy.com/> accessed 28 March 2023.

53 <http://www.billybot.co.uk/> accessed 28 March 2023.

54 <https://blog.rossintelligence.com> accessed 28 March 2023.

55 <https://donotpay.com/> accessed 28 March 2023.

56 <https://robotlawyerlisa.com/> accessed 28 March 2023.

simpler tasks. They can be used both for the purposes of civil and criminal proceedings. Finally, these programmes will also assist law enforcement and the judiciary themselves not only as assistive tools for search activities, but also for the translation of documents in proceedings.

3.1.2 Image recognition

Another significant area of AI use is computer vision and image recognition. These terms are sometimes used interchangeably, but image recognition encompasses a range of techniques and processing activities, such as image identification, or image classification based on computer vision.[57]

One of the measures used for crime detection and prevention is face analysis and face recognition. An AI algorithm can simultaneously not only detect and recognise a face, but also assess its layout, position, determine gender, age, emotions and so on. Face analysis based on ANNs, machine learning and computer vision allows image and video footage to be analysed to verify identity, background, health and even possible intentions. The systems can accurately recognise a face even if it has undergone some kind of transformation for example due to the ageing process. These systems may in part be intended for commercial use, but they can also be used for crime prevention and detection in law enforcement and evidentiary purposes.[58]

AI systems will not be limited to face analysis alone - they can also include the analysis of objects, objects, shapes, patterns, etc. contained in photographs or video footage. These systems can also be used for security maintenance, detection of dangerous objects, such as in the case of security tools used at airports.[59] The image recognition and other computer vision tools can be also successfully applied for the purposes of evidence processing.

3.1.3 Predictive analysis

The overall aim of AI predictive analytics is to try to predict future outcomes based on pre-collected data and statistical modelling, data mining and machine learning. This analysis will rely on significant data resources (big data), for its practical use. In order to make any predictive estimates, the AI system will first need to process a range of information collected in various types of databases.

57 Poonam Yadav, Hukum Singh and Kavita Khanna, 'Computer Vision, Its Applications, and Challenges' (22 February 2022) Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022 available at <https://papers.ssrn.com/abstract=4041050> accessed 20 April 2023.

58 Antonio J Colmenarez and Thomas S Huang, 'Face Detection and Recognition' in Harry Wechsler and others (eds), *Face Recognition: From Theory to Applications* (Springer 1998).

59 Guadenz Boesch, 'Image Recognition: The Basics and Use Cases (2023 Guide)' (*viso.ai*, 10 December 2023) <https://viso.ai/computer-vision/image-recognition/> accessed 30 May 2024.

Then it needs to identify recurring patterns, relationships, or trends, and derive specific conclusions from this processing.[60] Predictive AI will be applicable both for use by lawyers – eg to try to predict the outcome of a case based on historically collected data, but it will also be an essential tool for authorities, especially the police and activities aimed at individual and general prevention and detection of crime.[61]

Some of the solutions have already been put into practice. Such systems have been implemented to predict the type of crime, its date, time and place of its commission. Examples in this regard include Risk Terrain Modeling,[62] PredPol,[63] Italian X-Law,[64] KeyCrime,[65] German Precobs also used in Switzerland,[66] or British HART.[67] These systems can be used for general prevention, focusing on location-based tools, but can also be used to detect perpetrators of previously committed crimes.[68] In the case of investigative activities and real-time responses, ie taken in the time frame between deciding to commit a crime and committing it, they can also influence the final implementation of intent.[69] Individual-oriented systems such as COMPAS for example, have also come into use.[70] Systems that support the work of law enforcement agencies, such as facial recognition systems, DNA analysis, gunshot detection are also emerging with the use of artificial intelligence.[71]

60 Maciej Wach and Iwona Chomiak-Orsa, 'The Application of Predictive Analysis in Decision-Making Processes on the Example of Mining Company's Investment Projects' (2021) 192 Procedia Computer Science 5058.

61 Serena Quattrocolo, *Artificial Intelligence, Computational Modelling and Criminal Proceedings: A Framework for A European Legal Discussion*, vol 4 (Springer International Publishing 2020) 39–40.

62 <https://www.riskterrainmodeling.com/> accessed 28 March 2023.

63 <https://www.predpol.com/> accessed 28 March 2023.

64 <https://www.xlaw.it/presentazione/index_eng.asp> accessed 28 March 2023.

65 <https://keycrime.com/> accessed 28 March 2023.

66 <https://www.stadt-zuerich.ch/portal/de/index/politik_u_recht/stadtrat/weitere-politikfelder/smartcity/english/projects/precobs.html>, <https://land-der-ideen.de/en/project/precobs-software-for-predicting-crimes-355> accessed 28 March 2023.

67 Marion Oswald and others, 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality' (2018) 27 Information & Communications Technology Law 223–250.

68 Quattrocolo (n 61) 39–40, Manuel A Utset (n 17), 163–183.

69 Utset (n 17) 173.

70 Danielle Kehl, Priscilla Guo and Samuel Kessler, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' (*Berkman Klein Center for Internet & Society, Harvard Law School*, 25 August 2017) <https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf> accessed 22 March 2023.

71 Nipuni A Wickramarathna and Eata Edirisuriya, 'Artificial Intelligence in the Criminal Justice System: A Literature Review and a Survey' (*General Sir John Kotelawala Defence University*, 2022) <http://ir.kdu.ac.lk/handle/345/5244> accessed 30 May 2024, 4–5.

3.2 Risk of bias and malfunction

The examples given above are only a short summary of the possibilities of AI application within the criminal proceedings. The scope of tools available for use by both State actors and private entities, including lawyers and parties to the proceedings, is far wider. However, one of the major concerns that make this theoretical argumentation valid in a practical dimension is the fact, that the AI systems have the proven track record of bias. The most concerns have been raised in relation to racial discrimination[72], but there are other issues connected eg with gender, sexual orientation or even disabilities.[73] The face recognition systems (including Microsoft, IBM, Megvii) have been proven of less accuracy in case of women and other ethnicities and races than Caucasian.[74] The systems can therefore appear imprecise and affect the final result of the proceedings, causing serious consequences for the people affected.[75] The natural language processing systems used eg in the translation and search-oriented tools, also reflect the bias and can contribute to perpetuation of the stereotypes.

Many controversies are connected with profiling systems. They have reflected racial, socioeconomical and gender bias. In case of gender, there can appear

72 Zo Ahmed, Bertie Vidgen and Scott A Hale, 'Tackling Racial Bias in Automated Online Hate Detection: Towards Fair and Accurate Detection of Hateful Users with Geometric Deep Learning' (2022) 11 EPJ Data Science 1, Will Douglas Heaven 'Predictive Policing Algorithms Are Racist. They Need to Be Dismantled' (*MIT Technology Review* 17 July 2020) <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/> accessed 28 March 2023, 'Test Algorithms for Bias to Avoid Discrimination' (*European Union Agency for Fundamental Rights*, 30 November 2022) <http://fra.europa.eu/en/news/2022/test-algorithms-bias-avoid-discrimination> accessed 28 March 2023.

73 Lydia X Z Brown and others 'Ableism And Disability Discrimination In New Surveillance Technologies: How New Surveillance Technologies in Education, Policing, Health Care, and the Workplace Disproportionately Harm Disabled People' (*Center for Democracy and Technology*, 24 May 2022) <https://cdt.org/insights/ableism-and-disability-discrimination-in-new-surveillance-technologies-how-new-surveillance-technologies-in-education-policing-health-care-and-the-workplace-disproportionately-harm-disabled-people/> accessed 28 March 2023, European Network of National Human Rights Institutions (ENNHRI), 'Artificial Intelligence and Gender Equality: How Can We Make AI a Force for Inclusion, Instead of Division?' (*ENNHRI*, 8 March 2023) <https://ennhri.org/news-and-blog/artificial-intelligence-and-gender-equality-how-can-we-make-ai-a-force-for-inclusion-instead-of-division/> accessed 28 March 2023, Fabian Lütz, 'Gender Equality and Artificial Intelligence in Europe. Addressing Direct and Indirect Impacts of Algorithms on Gender-Based Discrimination' (2022) 23 ERA Forum 33.

74 The example can be a New Zealand case in which an Asian person was denied a passport due to 'closed eyes'. Can Yavuz, 'Machine Bias Artificial Intelligence and Discrimination' (Master thesis, University of Lund 2019), 'The Perpetual Line-Up' (*Perpetual Line Up*) <https://www.perpetuallineup.org/> accessed 28 March 2023.

75 Clare Garvie, Jonathan Frankle and Alvaro Bedoya 'Unregulated Police Face Recognition in America - Perpetual Line Up' (*GeorgeTown Law Center on Privacy & Technology*, 18 October 2016) <https://www.perpetuallineup.org/> accessed 28 March 2023.

the problem of so-called reverse discrimination, as the existing statistics are mainly focused on men.[76] There have been cases, where profiling overly targeted transgender or disabled people,[77] but mostly there have been severe cases of racial bias. The systems, such as COMPAS and other programs dedicated for individual prevention were introduced eg in Los Angeles and Chicago. Chicago launched a program, which relied on data from arrest records and social media to assess risk of participating in shootings. The result of this program was so-called "Strategic Subject List" that was meant to facilitate the work for the police. Due to controversies, the Programme was officially closed in 2020.[78]

Generally speaking, the source of the problem is that machine learning models reflect the biases of those designing the systems, implementing them, and collecting certain data. The human influence cannot be separated from the collected data. This human factor can, especially in the light of the process of data training, influence the final decision making. Finally, even if there is no bias in the collected data, the system can simply produce an error.[79] As a result, it is necessary, that the court, but also the defendant and his lawyer can obtain the information on how the decision was made by the AI systems and why such recommendation was proposed.

3.3 Impact on defence rights

With such a broad scope of actions, AI systems can significantly impact the course of proceedings and influence decisions of the prosecutor and the court. However, with the explainability issues in mind and due to general complexity of the systems, the particular defence rights can appear ineffective. The use of new technologies in the criminal proceedings already raises serious controversies about the equality of the parties, especially in terms of surveillance.[80] The State in this case has many more measures, resources and information at its disposal. This disparity and doubts as to the effectiveness of particular rights are further deepened in the case of the implementation of AI systems into the proceedings.

ECHR in its art 6 divides general right of defence into particular rights. This division can be helpful to identify particular threats to the specific elements of

76 It can affect mostly other legal issues, connected with especially family law and granting the custody. Can Yavuz, 'Machine Bias Artificial Intelligence and Discrimination' (2019) <https://www.researchgate.net/publication/334721591> accessed 28 March 2023.

77 Lidia X Z Brown and others (n 73).

78 However, there still appear concerns that the Police uses it unofficially 'Stop and Frisk | ACLU of Illinois' (11 August 2017) <https://www.aclu-il.org/en/campaigns/stop-and-frisk> accessed 5 April 2023.

79 Sasanka Sekhar Chanda and Debarag Narayan Banerjee, 'Omission and Commission Errors Underlying AI Failures' [2022] AI & SOCIETY.

80 Radina Stoykova, 'Digital Evidence: Unaddressed Threats to Fairness and the Presumption of Innocence' (2021) 42 Computer Law & Security Review 105575.

defence rights. art 6(3) of the ECHR sets the minimum standard and the core of defence rights. It can be considered an optimum starting point for the analysis. The elements included in the article are:

'(a) to be informed promptly, in a language which he understands and in detail, of the nature and cause of the accusation against him;

(b) to have adequate time and facilities for the preparation of his defence;

(c) to defend himself in person or through legal assistance of his own choosing or, if he has not sufficient means to pay for legal assistance, to be given it free when the interests of justice so require;

(d) to examine or have examined witnesses against him and to obtain the attendance and

examination of witnesses on his behalf under the same conditions as witnesses against him;

(e) to have the free assistance of an interpreter if he cannot understand or speak the language used in court.'

Controversies with the AI use are mostly connected with the right to information and access to case files. The right to information is one of the basic rights of the defendant, which enables them to take action - both in response to the actions of the prosecution and on his own initiative. It provides means to ensure the active and effective participation of defence counsel in the proceedings as well. The right to information, broadly defined, encompasses not only the right to be informed about the nature and cause of the accusation and information about the charges, but also the right to be informed about the rights of the defendant.[81] Knowledge of procedural rights is an essential safeguard of the effective use of defence rights in the course of the trial, while preventing possible abuse of power by the law enforcement agencies. It thus facilitates the implementation of the principle of equality of arms and the adversarial model of the trial.[82]

The most important conclusion that can be drawn from ECtHR case law is that information must be clear, interpretable. The same may be applied to the case materials. According to ECtHR, the defendant should be able to draw clear conclusions from the case files, allowing them to conduct effective defence.[83] It should not be presented only as a formal fulfilment of an obligation, but should equip the defendant with the real tools to take action in the proceedings. Both information about the charges, their basis and evidence for their support should

81 European Court of Human Rights, 'Guide on Article 6 of the Convention – Right to a fair trial (criminal limb)' (29 February 2024), 34–35, 80–81.

82 *Pelissier and Sassi v France*, App no 25444/94 (ECtHR, 25 March 1999), para 54; *Dallos v Hungary*, App no 29082/95 (ECtHR 1 March 2001), para 47.

83 *Migoń v Poland*, App no 24244/94 (ECtHR, 25 June 2002); *Schops v Germany*, App no 25116/94 (ECtHR 13 February 2001); *Garcia Alva v Germany*, App no 23541/94 (ECtHR, 13 February 2001); *Lietzow v Germany*, App no 24479/94 (ECtHR, 13 February 2001).

be presented in such a way, that defendant can draw some useful conclusions for their case as such information will be useful to benefit from other rights enlisted in the art 6 ECHR.[84] So, if the defendant will be presented with the files but could not encode from them any real information, both the right to information and the right to have adequate time and facilities for the preparation of his defence will be put into question.

In case of defence in person, the defendant should be provided with an opportunity to effectively participate in the proceedings, and thus be equipped with a number of procedural rights that will guarantee them this effective participation. In addition to the already mentioned rights to information and access to case materials, the defendant also has the right to question the legitimacy of the procedural actions carried out, to present evidence in support of their claims, to file motions, submit evidence, and to question the decisions made during the proceedings.[85]

The defence counsel, just like the defendant, must be given the actual opportunity to take action in the proceedings. The mere formal granting of the right to legal assistance is not a sufficient solution in the light of the current standards of the right to access to lawyer. The right of access to defence counsel should not be illusory requirement with no practical use, but existing regulations should, once again, guarantee the real effectiveness of this right. In order to be able to actually fulfil their role, the lawyer must have certain procedural measures at their disposal to ensure that they can actively participate especially in the evidentiary proceedings and exercise control over the course of the trial.[86] The aim and essence of the access to the lawyer is the ability to actively represent the defendant by first learning the facts, then establishing a line of defence

84 Stefan Trechsel, 'The Right to Be Informed of the Accusation' in Stefan Trechsel and Sarah Summers (eds), *Human Rights in Criminal Proceedings* (OUP 2006).

85 Stefan Trechsel, 'The Right to Defend Oneself and to Have the Assistance of Counsel' in Stefan Trechsel and Sarah Summers (eds) (n 84), Arkadiusz Lach *Rzetelne postępowanie dowodowe w sprawach karnych w świetle orzecznictwa strasburskiego,* (Wolters Kluwer 2018*)* 125. *Lüdi v Switzerland* App no 12433/86 (ECtHR, 15 June 1992), paras 49-50; *Lucà v Italy* App no 33354/96 (ECtHR, 27 February 2001), para 39; *AAl-Khawaja and Tahery v UK* App no 2676/05 and 22228/06 (ECtHR, 15 December 2011), para118.

86 Thomas Barkhuysen and others, 'Right to a Fair Trial', in Pieter van Dijk and others (eds) *Theory and practice of the European Convention on Human Rights* (Cambridge 2018) 637.

in consultation, confidentially (what is an another crucial condition of the effective defence),[87] with the client, and finally conducting the defence at trial.[88]

Consequently, if the defendant cannot draw any conclusions from the case materials or if they will be burdened by the flood of incomprehensible information, they will not be able to take any real actions and challenge it effectively. The same effect can appear with the defence counsel participation. In case of a significant digitalization, lack of transparency in the use of particular systems, the result can be the same as if the defence counsel did not take part in the proceedings at all. If there is no knowledge on the systems used and its working cannot be traced, the lawyer cannot in fact initiate any type of review and execute effective control over the course of the trial. When tracking the algorithm of AI either demands the specialist knowledge, not accessible to the average citizen, or is even impossible, the purposes of the right to information cannot really be met. As a result, the defendant in fact cannot effectively challenge the procedural actions on which his potential conviction could be based. Moreover, it is highly possible that the judge in most cases will also not possess the broad technological knowledge. Therefore the ability to assess the eventual malfunction, bias in the system or the disproportionality of the measure and in this regard refer to the appeal also should be put into question.[89]

Similar problem arises with the right of translation. The right to translation has a significant impact on all defence rights. It is difficult to assume that the defendant can consciously and fully exercise his information rights or the opportunity to actively participate in the proceedings if they does not understand the content of proper instructions, charges or evidence gathered during the proceedings. It is crucial to the fairness of the proceedings and the effectiveness of the right of defence that the translation is of an adequate quality. The competent authorities (the court or the prosecutor) should watch over this quality, also through *ex post* reviews.[90] The parties to the proceedings should be able to challenge quality of the translation as well.[91] The defence should have

87 *S. v Switzerland*, App no 12629/87 and 13965/88 (ECtHR, 28 November 1991) and art 4 of the Preamble of Directive 2013/48/UE, Commission, 'Report on the implementation of the 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' COM(2019) 560 final.

88 Paweł Wiliński, *Rzetelny proces karny w orzecznictwie sądów polskich i międzynarodowych* (Wolters Kluwer 2009) 26-27.

89 See also Luciano Cavalcante Siebert and others, 'Meaningful Human Control: Actionable Properties for AI System Development' (2023) 3 AI and Ethics 241.

90 *Kamasinski v Austria*, App no 9783/82 (ECtHR 19 December 1989), *Knox v Italy* App no 76577/13 (ECtHR, 24 January 2019), *Hermi v Italy*, App no 18114/02 (ECtHR, 18 October 2006).

91 *Kamasinski v Austria* (n 91).

sufficient means and opportunities to initiate review over translation or the interpretation. Therefore, there should be a clear possibility to analyse and recreate the mechanism of the system. If there is not – once again the effectiveness of the rights can be questioned.

Another major controversy, this time connected with the use of AI–based systems by the defence, appears due to the dynamic development of the systems created for the purposes of legal aid.[92] These systems can be prone to malfunction as well and can constitute cause for deterioration in the situation of the defendant. Existing case law of ECtHR addresses the cases when the lawyer does not perform their duties in a diligent and active way. ECtHR sees these situations as a lack of effective defence. If the system, on which the defence was based, was in fact faulty, the result would be in fact the same. The question appears, what can be done to limit the possible threats and disadvantages, mitigate the negative effects, and finally if the systems should be subjected to any type of review.

Besides the elements included in art 6(3), the use of AI systems and lack of explainability can influence other elements of art 6 that are in fact strictly connected with effective execution of defence rights. One of the most prominent examples is reasoning of judicial decisions. According to case law of ECtHR, judgments of courts should adequately state the reasons on which they are based.[93] They should prove to the parties that the decision was based on objective arguments and demonstrate that they were heard. The reasoning of the court is also necessary to preserve defence rights.[94] Without it, it is impossible to fully exercise any available right of appeal.[95] When the defendant does not possess sufficient knowledge on the reasons that led the court to its decision, they cannot really challenge it and address it.

Lacking explainability can in fact cause serious deficiencies in reasoning of judicial decisions. When we take into consideration AI systems that assist in rendering a judgment, due to black box, the possibilities for presenting reasons on which the decision is based can be very much limited. The defendant may receive simply the result of the systems' workings, but not the real grounds in the commonly accepted sense. It is also connected with the limited abilities of the judge that is making use of AI systems. The possibilities for the defendant

92 Daniel Martin Katz and others, 'GPT-4 Passes the Bar Exam' (15 March 2023) 382 Philosophical Transactions of the Royal Society A (2024), https://doi.org/10.1098/rsta.2023.0254 accessed 30 May 2024, <https://openai.com/blog/chatgpt> accessed 5 April 2023, <https://firmsy.com/> accessed 5 April 2023, <http://www.billybot.co.uk/> accessed 5 April 2023, <https://blog.rossintelligence.com> accessed 5 April 2023, <https://donotpay.com/> accessed 5 April 2023, <https://robotlawyerlisa.com/> accessed 5 April 2023.

93 *Moreira Ferreira v Portugal*, App no 19867/12 (ECtHR, 5 July 2017), para 84.

94 See 'Guide on Article 6' (n 81).

95 *Hadjianastassiou v Greece*, App no 12945/87 (ECtHR, 16 December 1992).

to challenge such decisions are in this case restricted and can affect the successful conduct of defence.

All these deficiencies combined can as a consequence influence the overall fairness of the proceedings. The ineffective rights cannot lead to a fair trial, which should be based on equality of arms and effective participation. According to case law of ECtHR, art 6 as a whole guarantees the right of defendant to participate effectively in a criminal trial.[96] It also includes a right to follow the proceedings.[97] Defence rights in their essence should contribute to implementation of these principles. It is a fundamental guarantee of the truly effective realization of the right to a trial. It grants the defendant the opportunity to choose their behaviour during the trial within the limits of the law, with the possibility of passive behaviour which is reflected in a privilege against self-incrimination, but also active participation in the proceedings and in taking steps to prove certain circumstances and present their version of events.[98] Without an adequate degree of explainability, this effective participation can be substantially restricted. Therefore, the proper regulation is clearly needed to safeguard the essence of right of defence.

3.4 Prospects for regulation

The need for imposing specific obligations on the public sector with regard to the AI use and explainability has been raised before, especially by the Council of Europe. The CAHAI (Ad hoc Committee on Artificial Intelligence) called for a legally binding transversal instrument that would include a series of provisions on legal safeguards applicable do AI systems used for the purpose of deciding or informing decisions impacting "the legal rights and other significant interests of individuals and legal persons".[99] These safeguards should, at least, include the following: the right to an effective remedy before a national authority (including judicial authorities) against such decisions, the right to be informed about the application of an AI system in the decision-making process, and the right to choose interaction with a human in addition to or instead of an AI system, and the right to know that one is interacting with an AI system rather than with a human. The systems used should be trustworthy, ie intelligible, traceable and auditable. One general conclusion from these recommendations when applied to criminal proceedings is that the defendant should be able to take real actions and should be equipped with effective tools to challenge the AI-based evidence or decision process. To do so, they should be able to obtain all the relevant information, then understand it, and finally take proper measures.

96 *Murtazaliyeva v Russia*, App no 36658/05 (ECtHR, 18 December 2018).

97 *Moreira Ferreira v Portugal* (n 94).

98 Stefan Trechsel, 'The Right to Defend Oneself and to Have the Assistance of Counsel' in Stefan Trechsel and Sarah Summers (eds), *Human Rights in Criminal Proceedings* (OUP 2006).

99 CAHAI (n 41).

As for the scope of possible action for the defendant, there can be significant differences depending on the applicable model of the proceedings. In the adversarial models of the trial the defendant is equipped with a broader right to submit evidence gathered by private individuals, but must rely on his own or his lawyers' abilities in this regard. If the model is more inquisitorial, the evidence most likely will be obtained by the prosecutor and taken by the court. But in this case the question arises, if the defendant can effectively demand such information about the AI systems used. It has been emphasised before, that due to the fact that law enforcement agencies have a wide range of resources and the entire State apparatus at their disposal, the data acquisition rights are often one-sided in the proceedings where the new technologies are used. In such cases, there has been a visible trend of pre-trial phase domination in evidence collection. Also, due to the possibility of collecting data in a large amount and on a large scale, this leads to a situation, where defendant de facto has to prove his innocence.[100] This can be even more evident if the AI systems are used, as they can cause so much transparency issues that the defendant is not only burdened by them, but is not able to challenge them effectively.

Regardless of the model of the proceedings, defendant must be equipped with the effective means to obtain the data on the AI systems applied in the proceedings. Some interesting solutions has been introduced in the Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (Proposal for the AI Liability Directive).[101]

According to the proposed Directive, Member States should ensure that national courts are empowered, upon the request of a claimant to order the disclosure of relevant evidence about a specific high-risk AI system that is suspected of having caused damage from a provider or a user. It is a significant facilitation for the claimant that gains access to the necessary tools for proving his statements. The Directive refers to the "evidence". However such a right to disclose the relevant information about AI systems itself could be attributed to the defendant. The ability to address the court and ask for the information on how the particular decision (both on the pre-trial and trial stage) was made with the reference to the inner workings of the system, could improve significantly the situation of the defendant.

If the ability to gather the information on the AI systems is guaranteed, then adequate measures should be introduced to guarantee the practical understanding of the files and effectiveness of the defence rights. However, different groups and entities will need different types of explanations.[102] The needs of

100 Radina Stoykova, 'Digital Evidence: Unaddressed Threats to Fairness and the Presumption of Innocence' (2021) 42 Computer Law & Security Review 105575.

101 Commission, 'Proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence' COM (2022) 496 final (hereinafter 'Proposal for the AI Liability Directive').

102 See also The Royal Society, *Explainable AI: the basics Policy briefing* (The Royal Society 2019).

the court, huge legal companies and defendant acting on its own may not be the same. The assistance of experts should also be guaranteed in the proceedings – both for the use of the defence and the court. The defendant should be able to refer to the expert knowledge to effectively challenge the gathered evidence in case of any objections as to their reliability. The court should make use of such assistance of experts in case of any doubts connected with the evidence submitted by the parties to the proceedings, including the prosecution and suspected malfunctions in the systems.

As for the explainability models described above, the decompositional approach would be of great benefit for the experts. The use of eg surrogate models could assist the analysis and help really understand how the system reached its decision. Then, the processed information could be presented to the court and defence. Another way can be the strive for increased traceability. This is for example achievable by limiting the ways through which system decisions can be reached and creating a narrower scope for machine learning rules and features. This can reduce the problems connected with neural network hidden layers. It should represent more broadly the connections between individual neurons and the emerging internal dependencies between them. The example can be DeepLIFT (Deep Learning Important FeaTures), which identifies the activation of each neuron and its counterpart, thus allowing the reconstruction of the connection chain.[103]

The decompositional approach is however of little importance for the defendant themselves. With no knowledge on the technology and the systems used, the defendant still does not understand to any extent, how the particular decision has been made – especially in the case of receiving the judgment and the particular penalty. If the AI systems are used on a large scale in the proceedings, it can result in decrease of general trust in the state. The solution can be the local interpretability (subject-centric approach). The defendant can gain insight into the grounds for particular decision, search for possible bias and draw the attention of the court to it.

Finally, in case of the use of AI systems for the benefit of the defendant, eg for the purpose of translation or to shape defensive strategies, the systems definitely should also be subjected to the review. Two situations can be differentiated – if the defendant uses the system on his own (especially not the free, publicly available one, but purchased) and if the system is provided to any extent by the state. In the latter case, the systems should unquestionably

103 Jon Rueda and others, "'Just" Accuracy? Procedural Fairness Demands Explainability in AI-Based Medical Resource Allocations' [2022] AI & SOCIETY, Bas HM van der Velden and others, 'Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis' (2022) 79 Medical Image Analysis 102470, Avanti Shrikumar, Peyton Greenside and Anshul Kundaje, 'Learning Important Features Through Propagating Activation Differences' (*airXiv*, 12 October 2019) <http://arxiv.org/abs/1704.02685> accessed 26 April 2023.

undergo the introductory control process assessing their quality, with a view to the requirements of the AI Act. Moreover the defendant also should be able to initiate the review. If the AI system would be inefficient and malfunctioned, it should make an effective ground for an appeal. In the former, the systems also must be developed in accordance with the AI Act, however the potential consequences of the inadequacies and mistakes can be debatable. The defendant possibly could make use of the liability regime provided that the applicable conditions are met, but the ground for appeal is in this case very questionable as it can be perceived as the execution of free choice of the defendant.

4. Summary

Despite ongoing attempts to regulate the AI use in different spheres of legal and social reality, the AI systems' application in the criminal proceedings is still underregulated, particularly in the area of defence rights. The main aspect of the upcoming regulation should be safeguarding explainable AI as the black box is the biggest threat to the effective defence rights. The general need for xAI has been identified before - both in terms of public and private sector. However, it should turn into some specific obligations – both for the state and systems' developers.

In the current state of regulation the main threat is that the all the defence rights will not be effective. The AI systems' implementation into the proceedings does not result in the breach of any particular right *per se* at the first glance. They can be formally observed, but deeply ineffective. The decision made as a result of proceedings in which the information about the real nature of the evidence is in fact lacking, can be perceived as arbitrary and therefore contradict the aim of the AI use in the proceedings. Lacking information can further affect the ability to take real actions in the proceedings and issue any type of review. The review must be effective and accessible to the defence as there are considerable risks of bias and malfunction. They mostly have been raised in connection to the predictive policing, but can also affect other systems, for example ones based on NLP. Therefore to guarantee the effectiveness of defence rights, the explainable AI must be guaranteed. The explainability should be safeguarded by imposing obligations on the developers of the systems and implementing rights attributed to the defendant.

First of all, the legislator should ensure that providers of systems used in criminal proceedings strive for local interpretability and real explainability. The traceability of the AI –based systems should be ensured (to some extent introduced in the AI Act), but also means such as surrogate models should be disseminated. The general information about systems themselves does not give a proper extent of explainability sufficient for realization of the right to information. The main advantage of the local interpretability is that the defendant

themselves can gain insight into the reasons and premises for particular decision, which can help in retaining trust to the judiciary. The decompositional approach on the other hand and the use of surrogate models can be of great usefulness in cases where the review will be needed.

Then, the effective right to demand information must be introduced, possibly based on the Proposal for the AI Liability Directive example. The defendant should be given a clear right to challenge the results obtained by the use of AI systems with reference to the inner workings of the systems. To combine these two elements and guarantee the practical understanding of the files and effectiveness of the defence rights, the assistance of experts should also be guaranteed in the proceedings. The defence should be able to refer to such assistance and put AI-based evidence under their review and so should the court.

Finally, in case of the use of AI systems for the benefit of the defendant, the systems must undergo a proper review and should be labelled as high-risk. If the State is a provider of such system, the *a priori* review should be especially diligent. In case of any malfunction, it should make an effective ground for an appeal.

Some of the conclusions may appear overly future-oriented, however the rapid dissemination of the AI-based tools, such as ChatGPT, illustrates the possibilities for the AI systems create and can serve as a prognosis for the future. It is increasingly likely that the AI in some time will be commonly used in criminal proceedings. As a result, the legal act safeguarding the rights of the parties to the proceedings including the defendant should be issued. The one comprehensive regulatory act covering all spheres of application, such as AI Act, can be no longer perceived as sufficient in this case. In the criminal proceedings context, the most optimal and beneficial choice would be in this case a directive dedicated strictly to this sphere of AI use, as it would create a minimum standard for defence rights and at the same time it would leave some regulatory freedom to the Member States, allowing it to be adjusted to local needs and legal systems.

Determining the possible implementation of AI systems that can be compatible with procedural rights, including right of defence, is one of the key issues that judicial and law enforcement systems are going to face. It is also not without significance taking into account the field of European cooperation in criminal matters, mutual recognition of judgments and law enforcement cooperation. Significant system differences can significantly impede an effective response to criminal activities. Doubts concerning the systems assisting in rendering judicial decisions and lack of possibilities to review them can make it impossible to recognize such decisions in other EU Member States. As a result issuing such Directive could be beneficial for the international cooperation.

One final point to be made is connected with the issue of effectiveness versus the fair trial standards. The more complex the system is, the more concerns appear connected with transparency, explainability and, as a consequence, the

defence rights. It concerns eg unsupervised learning, AI systems based on artificial neural networks with hidden layers of nodes and deep learning. However, the more complex systems allow for higher efficiency and effectiveness. As a result, there is clearly a need for balancing the values – possible gain vs. human rights protection. The approach that the EU as well as Member States will choose – more guarantee or more gain-oriented will define the shape of regulation. As it can be drawn from the practice and recent statements of EU bodies, it appears that the EU has chosen the more protective approach and it will conduct harmonization based on the respect for human rights.[104] It can result in decrease in efficiency and slower development of the AI systems, but it will allow to safeguard the minimum standards arising from the provisions of the Convention, including fair trial standards and defence rights. Issuing the Directive with specific guarantees can prevent disregarding them by the particular States in favour of increased effectiveness and acceleration of the proceedings.

104 Giovanni De Gregorio, in *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (CUP 2022), ch. 7 'The Road Ahead of European Digital Constitutionalism'.