Gatt L, Caggiano IA, Gaeta MC, Troisi E, Savella R, Pratesi F, Trasarti R, 'FRIA implementation model according to the AI Act' (2024) 2 EJPLT.
DOI: https://doi.org/10.57230/EJPLT242LGIACMCGETRSRTFP

EJPLT
ISSN 2704-8012

FRIA implementation model according to the AI Act

Modello di attuazione FRIA in applicazione dell'AI Act

Lucilla Gatt
Full professor of Private Law, Director of the Research Centre in European Private Law (ReCEPL) Suor Orsola Benincasa University of Naples

Ilaria Amelia Caggiano
Full professor of Private Law, Vice-Director of the Research Centre in European Private Law (ReCEPL)
Suor Orsola Benincasa University of Naples

Maria Cristina Gaeta
Lecturer in Private Law, Ph.D., Scientific Secretary of the Research Centre in European Private Law (ReCEPL)
Suor Orsola Benincasa University of Naples

Emiliano Troisi
Ph.D., Junior Researcher Research Centre in European Private Law (ReCEPL),
Suor Orsola Benincasa University of Naples

Roberta Savella
Research Fellow Project SoBigData++
Institute of Information Science and Technologies "Alessandro Faedo" (ISTI),
National Research Council

Francesca Pratesi
Researcher Project SoBigData++
Institute of Information Science and Technologies "Alessandro Faedo" (ISTI),
National Research Council

Roberto Trasarti
Researcher Project SoBigData++
Institute of Information Science and Technologies "Alessandro Faedo" (ISTI),
National Research Council

Abstract
The paper presents the FRIA project aimed at researching and specifying a methodology to assess the impact of Artificial Intelligence (AI) systems on fundamental rights, as recognised by the international and European regulations of hard law and soft law, with the specific reference to the judicial sector as the field of analysis. The research methodology starts from the study of the existing legal and ethical frameworks concerning AI and human rights and the translation of the identified rules and principles into a set of synthetic requirements to create an automated risk assessment methodology. The research output is a prototype tool to support and automate the fundamental rights impact assessment of high-risk AI systems, which is in line with the requirements of the European Artificial Intelligence Act.

* The paper is the result of the work of a selected team of researchers of the Research Centre in European Private Law (ReCEPL), coordinated by the ReCEPL Director Prof. Lucilla Gatt and ReCEPL Vicedirector Prtof.ssa Ilaria A. Caggiano, as well as the National Research Council of Italy, coordinated by CNR First Researcher Doc. Roberto Trasarti. This work is supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu), and NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.

Abstract
Il lavoro presenta il progetto FRIA finalizzato alla ricerca e alla specificazione di una metodologia di valutazione dell'impatto dei sistemi di Intelligenza Artificiale (IA) sui diritti fondamentali, così come riconosciuti dalle normative internazionali ed europee di hard law *e* soft law*, con specifico riferimento al settore giudiziario come ambito di analisi. La metodologia di ricerca parte dallo studio dei quadri giuridici ed etici esistenti in materia di IA e diritti umani e dalla traduzione delle norme e dei principi identificati in un insieme di requisiti sintetici per creare una metodologia di valutazione automatica del rischio. Il risultato della ricerca è un prototipo di strumento per supportare e automatizzare la valutazione dell'impatto sui diritti fondamentali dei sistemi di IA ad alto rischio, in linea con i requisiti del Regolamento europeo sull'intelligenza artificiale.*

Keywords: AI Act; high-risk AI systems; impact assessment; FRIA; metrics methodology; ELS; risk management; Law&Ethics

Summary:

1. Introduction and background.*

In the AI regulation at a European and international level, the idea of ex ante protection as the main protection in relation to high-risk AI systems has been consolidated and a risk-based approach has been selected to measure the risk of damages resulting from the different AI applications.

The definition of AI system is currently provided by the Artificial Intelligence Act (Reg. 2024/1689/EU - AI Act), art. 3, para 1, no. 1, as a machine-based system designed to operate with different levels of autonomy and that may exhibit adaptiveness after deployment. Moreover, AI systems produce outputs (predictions, contents, recommendations or decisions) from the input received, that can influence physical or virtual environments. The AI Act classifies AI systems according to their possible risks for individuals, providing a specific regulation for high-risk AI systems (art. 6 ff. AI Act).

High-risk AI systems are defined as systems used as security components of a product or security products themselves, that are required to undergo a third-party conformity assessment (art. 6, para 1 AI Act). In addition, AI systems referred to in Annex III shall be considered to be high-risk, such as those related to the administration of justice listed at no. 8, Annex III, AI Act.

* The work was presented at the IEEE International Conference 'Metrology for eXtended Reality, Artificial Intelligence, and Neural Engineering' (MetroXRAINE), St Albans, London, October 21-23, 2024, and published in the IEEE MetroXRAINE 2024 Proceedings, 1224 – 1229.
The authors of this paragraph are Dr. Maria Cristina Gaeta and Dr. Emiliano Troisi.

In this light, the AI Act, at articles 9, para 2, and 27, in general terms refers to a Fundamental Rights Impact Assessment (FRIA) intended as a tool for an ex ante and interactive assessment of the possible risks that a certain AI system, classified as high risks, can pose, also with a view to managing these risks. It aims to prevent human rights violations and serve as an accountability tool for risk management system and impact assessment.

However, currently, no complete tools have been implemented with which to proceed to achieve this preventive protection. Indeed, there are mainly tools based on a self-assessment system, also provided by public entities, like the ALTAI (Assessment List for Trustworthy Artificial Intelligence)[1], a prototype of a tool to support AI R&D based on 7 ethical guidelines presented in 2020 by the High-Level Expert Group on Artificial Intelligence (AI HLEG). In the same direction, Council of Europe's European Commission for the Efficiency of Justice (CEPEJ) elaborated in 2023 the Assessment tool for the operationalisation of the European Ethical Charter on the use of artificial intelligence in judicial systems and their environment which consists of 29 questions. Furthermore, the CEPEJ Assessment Tool seems to be complementary to the Human Rights, Democracy, and the Rule of Law Impact Assessment (HUDERIA)[2], which is currently being designed by the Council of Europe's Ad Hoc Committee on AI (CAHAI), as uniform model for human rights impact assessments (HRIAs) of AI systems meeting the parameters and supporting the application of the recent approved Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law of 2024.

At the national level, instead, could be mentioned the FRAIA tool, elaborated by the Government of the Netherlands to help in mapping the risks to human rights in the use of algorithms and to take measures to address these risks[3]. These systems do not guarantee an overall view of fundamental rights, but rather take only certain aspects into account, lacking a comprehensive analysis necessary to ensure effective protection of the individuals, especially in risk-specific deployment contexts. Because of the lack of an effective tool suitable for the purpose of assessing the impact of AI on fundamental rights in a quantifiable, automatic and objective way, evaluation errors may occur which can then result in excessive or unexpected costs for individuals and companies.

Starting from the consequences of AI on human rights [4] and the introduction of the FRIA regulation in the AI Act, this study comes from the

[1] High-Level Expert Group on Artificial Intelligence, Assessment List for Trustworthy Artificial Intelligence (ALTAI), 17 of July 2020, freely available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[2] Council of Europe's Ad Hoc Committee on AI (CAHAI), Human Rights, Democracy, and Rule of Law Impact Assessment (HUDERIA).

[3] Government of the Netherlands, FRAIA (Fundamental Rights and Algorithms Impact Assessment), more info online: https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms

[4] A Quintavalla, J Temperman (eds.), Artificial Intelligence and Human Rights (Oxford University Press 2023); J Fjeld, N Achten, H Hilligoss, AC Nagy, M Srikumar, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI* (January 15, 2020) Berkman Klein Center Research Publication No. 2020-1, DOI: http://dx.doi.org/10.2139/ssrn.3518482; A Kriebitz, C Lutge, 'Artificial Intelligence and Human Rights: A Business Ethical Assessment, in Business and Human Rights Journal' (2020) Vol. 5, Issue 1, pp. 84-104; G De Gregorio, 'From Constitutional Freedoms to the Power of the Platforms: Protecting Fundamental Rights Online in the Algorithmic Society' (2019) in 11(2) *European*

identification of a gap between the theoretical affirmation of a rule and its practical application. In fact, the AI Act alone is not sufficient to specify how to implement a FRIA effectively, because there are no indicated parameters for the implementation of adequate measurement paths. This need is starting to become increasingly clear in the literature[5].

2. Methodology.*

The methodology of the research is based on two phases: (i) the study of the existing legal and ethical frameworks concerning AI and human rights; and (ii) the translation of the identified rules and principles into a set of synthetic requirements to create an automated risk assessment prototype as output of the research.

The measurement approach by score has a profound impact on the civil system for the protection of harm to human beings and requires a revision of the traditional law methodologies as well as real changes at a regulatory level. The project has investigated in both directions: (i) the effective protection of human beings, and (ii) the rethinking of the regulatory methods in order to strengthen the *ex ante* evaluation models for AI. As a matter of fact, high-risk AI has legal relevance and can be involved itself in the decision-making processes related to the design and placing on the market of intelligent products.

The research is based on a multidisciplinary approach of interaction between humanities (law) and technical sciences (engineers). In fact, legals are able to identify the reference parameters of the research (in this specific case, fundamental rights to be analysed to verify the impact that AI has on them) as well as quantify the type and levels of violation for the purposes of measurement, to be evaluated according to a judgment of balancing opposing interests. Engineers, on the other hand, have the task of developing the prototype for measuring the impact of AI and automating it.

To better understand the effective protection of human beings in relation to AI systems and to define the proper parameters to be applied for AI risk assessment, the project will start with the analysis of the current legal framework.

Journal of Legal Studies, pp. 65 ff.; C Fleissner, 'Inclusive Capitalism Based on Binary Economics and Positive International Human Rights in the Age of Artificial Intelligence' (2018) Vol. 17, Issue 1 Washington University Global Studies Law Review, pp. 201 ff.; F Raso, H Hilligoss, V Krishnamurthy, C Bavitz, K Levin Yerin, *Artificial Intelligence & Human Rights: Opportunities & Risks* (September 25, 2018) Berkman Klein Center Research Publication No. 2018-6, Available at http://dx.doi.org/10.2139/ssrn.3259344.

[5] A Mantelero, 'The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template' (March 30, 2024), DOI: http://dx.doi.org/10.2139/ssrn.4782126.

* The authors of this paragraph is Dr. Maria Cristina Gaeta.

3. State of the art.*

The state of the art on the legal and ethical regulation of AI is now quite complex. For further information, please refer to other ReCEPL studies on the topic[6]. In this paper, the focus is on the impact of AI on human rights.

3.1 International and European existing legal and ethical framework concerning AI and human rights.

The state of the art concerning the impact of AI systems on human rights starts from the analysis of the International and European fundamental rights charters. Particularly relevant are the Charter of Fundamental Rights of the European Union (FREU) of 2000, the Universal Declaration of Human Rights (UDHR) of 1948 and the European Convention on Human Rights (ECHR) of 1950.

With specific regard to the judicial sector at the International Level, then, the European Ethics Charter on the Use of Artificial Intelligence in Justice Systems and their Environment (CEPEJ Charter) of 2018 should be taken as a point of reference. The CEPEJ Charter sets out five key principles that should be respected in the design and use of artificial intelligence (AI) by judicial professionals: (1) respect for fundamental rights, (2) non-discrimination, (3) quality and security, (4) transparency, impartiality and fairness, and (5) user control. After the CEPEJ Ethics Charter, of great relevance, is the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law of 2024, which formulates fundamental principles and rules to be respected within the lifecycle of AI systems to comply with certain fundamental principles: human dignity and individual autonomy, equality and non-discrimination, as well as, privacy and personal data protection.

Concerning the EU level, the most recent hard law regulation on AI is the AI Act, as part of a wider package of policy measures to support the development of trustworthy AI, which also includes the AI Innovation Package[7] and the Coordinated Plan on AI[8]. Together, these measures aim to guarantee the safety of AI technologies and the protection of the fundamental rights of individuals and businesses in relation to AI applications. They will also strengthen uptake, investment and innovation in AI across the EU. Also from an ethical point of

* The authors of this paragraph is Dr. Maria Cristina Gaeta.

[6] MC Gaeta, L Aulino, E Troisi E (2023), 'The possible relationships between law and ethics in the context of artificial intelligence regulation' (2023) 44 Humana.Mente Journal of Philosophical Studies, 44, pp. 163 ff.; L Gatt L, IA Caggiano, MC Gaeta, L Aulino, E Troisi, 'The possible relationships between law and ethics applied to AI' (2023) 2 EJPLT, 111 ff., doi https://doi.org/10.57230/EJPLT232GCG.

[7] Communication from the European Commission on boosting startups and innovation in trustworthy artificial intelligence, 24 January 2024 COM(2024) 28 final, Official web page European Commission, AI innovation package to support Artificial Intelligence startups and SMEs, available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_383.

[8] Communication from the European Commission, Coordinated plan on Artificial Intelligence, 7 December 2018, COM(2018) 795 final, Official web page European Commission, Coordinated Plan on Artificial Intelligence, available at: https://digital-strategy.ec.europa.eu/en/policies/plan-ai.

view, the need to verify the impact of AI on individuals and on society in general [9], as well as on society, human psychology, financial systems, environment and trust, has been highlighted in European studies[10] and soft law acts[11].

With specific reference to the methodology of the risk-based approach and, more specifically, the risk impact assessment, it seems appropriate to highlight that, in the last years, the EU approach has been based on the analysis of the impact of the risk of a product/service on human being's rights, from the design phase up to the introduction on the market and its related use. The first example is that represented by the Data Protection Impact Assessment - DPIA (Artt. 35-36 Reg. 2016/679/UE - GDPR), then also followed by the risk assessment on human rights arising from the design and operation of the very large online platforms and very large online search engines (Art. 34 Reg. 2022/2065/UE – DSA), with explicit reference also to fundamental rights (art. 34, para 1, lett. b)[12]. This risk analysis concretely implies a measurement of the risks on the basis of a given scale of values. It is therefore the evolution of the concept of measurement, also analysed from a legal point of view[13].

This is the context in which FRIA developed, expressly provided by the AI Act, for which certain high-risk AI systems shall be avalutated, at the preventive stage and monitored throughout thei entire lifecycle, throuth an assessment tool of the impact on fundamental rights that the use of such a system may produce[14].

3.2 Italian state of the art AI regulation, with specific reference to its impact on human rights.

At the Italian level, the state of the art starts from the analysis of the White Paper of the Agency for Digital Italy, entitled 'Artificial Intelligence at the

[9] European Parliament Study, Artificial intelligence: From ethics to policy, PE641.507, June 2020

[10] European Parliament Study, The ethics of artificial intelligence: issues and initiatives, PE 634.452, March 2020; European Parliament Study, European framework on ethical aspects of artificial intelligence, robotics and related technologies, PE 654.179 – September 2020.

[11] European Parliament Resolution, Framework of ethical aspects of artificial intelligence, robotics and related technologies, of 20 October 2020, PROV(2020)0275; HLEG-AI, the Assessment List for Trustworthy Artificial Intelligence (ALTAI), 17 July 2020; European Commission White Paper on Artificial Intelligence – A European Approach to Excellence and Trust, 19 February 2020, COM 2020/65 final; HLEG-AI, Policy and investment recommendation for trustworthy AI, June 26, 2019; Communication from the European Commission, Building Trust in Anthropocentric Artificial Intelligence, 8 April 2019, COM 2019/168 final; HLEG-AI, A definition on AI: Main capabilities and Disciplines, April 8, 2019; HLEG-AI, Ethical Guidelines for Trusted AI, April 8, 2019; Communication from the European Commission, Building Trust in Human Centric Artificial Intelligence, 8 April 2019, COM(2019) 168 final.

[12] A Mantelero, Putting the DSA into practice (Verfassungsbooks 2023), pp. 107 ff..

[13] L Gatt, IA Caggiano, MC Gaeta, AA Mollo, 'BCI devices and their legal compliance: A prototype tool for its evaluation and measurement' (2022) 1 EJPLT 301 ff. DOI: https://doi.org/10.57230/EJPLT221LGIACMCGAAM.

[14] S Bertaina, I Biganzoli, R Desiante, D Fontanella, N Inverardi, IG Penco, A Cosentini, 'Fundamental Rights and Artificial Intelligence Impact Assessment: A New Quantitative Methodology in the Upcoming Era of Ai Act' (Jan 18, 2024), DOI: http://dx.doi.org/10.2139/ssrn.4698609; A Mantelero, *Beyond Data. Human Rights, Ethical and Social Impact Assessment in AI* (Springer 2022); H Janssen, Ah Lee MS, Singh J, 'Practical fundamental rights impact assessments' (2022) 30 International Journal of Law and Information Technology, pp. 200 ff., https://doi.org/10.1093/ijlit/eaac018; A Mantelero, SM Esposito, 'An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems' (2021) Computer Law & Securety Review, available at: https://doi.org/10.1016/j.clsr.2021.105561.

service of citizens', published in 2018. In particular, challenge 8 was focused on measuring the impact of Artificial Intelligence solutions in public administration[15].

The White Paper approached the issue from two points of view: that of the citizens (quality of life of people and customer satisfaction) and that of the institutions (optimisation of organisational processes). The White paper underlines the need to assess the impact of AI technologies by applying a multidisciplinary approach to the measurement with the aim of designing and developing trustworthy AI, ensuring reliability and transparency as well as reducing the risk of errors (and therefore possible damages for the individuals).

In 2024, the concept of 'impact', albeit in a broader sense, is taken up, in AgID's Italian Strategy for Artificial Intelligence 2024-2026[16], with reference to scientific research, public administration, companies, and education. In particular, the need to enhance applied AI research emerges, through initiatives co-designed by public-private partnerships, focusing on contexts of greater economic and social value for Italy and with a greater impact on citizens' well-being. In addition, the strategy points out that public administration processes need to be made more efficient, with a specific focus on clearly delineated areas of intervention by the PA also considering the impact and risks of the AI systems.

Furthermore, in May 2024, the Italian Council of Ministers approved the Italian bill on AI, in line with the AI Act, which regulates the principles of research experimentation, development, adoption and application of AI systems, promoting their proper, transparent and responsible use, in an anthropocentric dimension, as well as ensuring supervision on economic and social risks and, specifically, on the impact of AI on fundamental rights (Art. 1). It also introduces principled rules and industry provisions to promote the use of new technologies for the improvement of citizens' living conditions and social cohesion, as well as to provide risk management solutions based on an anthropocentric view. The bill provisions intervene in five specific areas: national strategy, national authorities, promotion actions, copyright protection, and criminal sanctions.

4. Cyberjustice and FRIA: Integrating AI in Judicial Contexts.*

The implementation of a risk assessment model for AI systems aimed at ensuring effective protection of fundamental human rights necessitates a context-specific approach. The risk analysis methodology must identify the extent to which a high-risk system is likely to pose significant risks to human rights, especially considering the sectorial application context. When possible (or expressly required; cf. art. 27 AI Act), it should also take into account the specific deployment environment, the stakeholders likely to be affected, and

[15] White Paper of the Artificial Intelligence Task Force of the Agency for Digital Italy (AgID), 'Artificial Intelligence at the service of citizens', Version 1.0, March 2018, Challenge 8.
[16] Agency for Digital Italy (AgID), 'Italian Strategy for Artificial Intelligence 2024-2026', April 2024, pp. 3-7.
* The authors of this paragraph is Dr. Emiliano Troisi.

their fundamental rights and freedoms actually exposed to risk from the system's use.

Our proposed model focuses on AI systems used in judicial and para-judicial (e.g., ADR/ODR) contexts for the resolution of legal issues. The risk assessment of AI must be contextualised to the specific area of use, which involves analysing the unique risks to fundamental rights associated with the intended use of AI systems in judicial contexts. The area of cyberjustice indeed presents a unique set of challenges and risks for the use of AI. These systems are not only subject to the AI Act (at least to the extent that they can be defined as high-risk), but also operate within a multi-level domain-specific regulatory framework. This framework includes supranational regulations, such as the recent Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024), and guidelines like the CEPEJ European Ethical Charter on the use of artificial intelligence in judicial systems and their environment (2018) and the related Assessment tool for the operationalization of the Ethical Charter (2023). Additionally, national rules often regulate how to implement or use AI in the justice sector in compliance with fundamental rights obligations (e.g., in Italy, the Bill providing for delegation to the government on Artificial Intelligence, 2024, currently under parliamentary discussion).

This comprehensive framework contributes to defining the limits and compliance requirements for cyberjustice systems and needs to be carefully considered, alongside the broader human rights framework and the current state of the art on ethical AI, in order to contextualise and detail the identification and evaluation of both the specific threats to fundamental rights that AI systems pose in the given context of application and deployment, and the appropriate measures for the overall mitigation and control of these risks.

Establishing an AI FRIA assessment tool that is specific to the judicial domain is indeed crucial for ensuring effective ex ante evaluation and iteractive governance of cyberjustice systems in a way that is relevant under the AI Act and in any case ensures proper assessments of AI systems helping to avoid unexpected or unwanted harms. Furthermore, it can also be used as a benchmark for conceptualising and defining specifications for the design of AI systems in the judicial domain.

5. Objectives.*

The main objective of the project is the identification of parameters for AI risk analysis and the creation of metrics for each parameter for the concrete measurement of AI risks for individuals. This main objective is a result of the application of the following procedural steps:

* The authors of this paragraph are Dr. Roberto Trasarti, Dr. Roberta Savella and Dr. Francesca Pratesi.

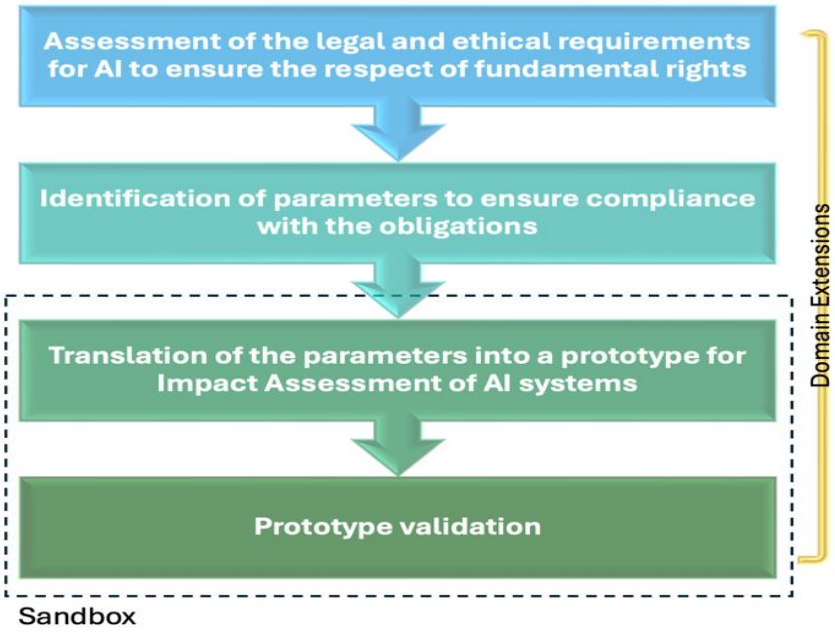

Image no. 1 – Objectives of the research

To reach this main objective, the output of the research consists in the design of tool prototype to support and automate the fundamental rights impact assessment of high-risk AI systems, in line with the requirements of the AI Act, i.e. an automated implementation model of the FRIA (Fundamental Rights Impact Assessment), required by the European AI Law, that has these characteristics.

6. The analytical process. *

The implementation model of the FRIA will be developed in the form of a semi-automatic prototype specifically designed for this purpose. The design phase of the project will focus on AI systems used in the administration of justice, as a relevant area of application of high-risk systems for which the AI Act requires deployers to perform a FRIA.

The research activity in this part of the project will start with two parallel assessments: (i) the selection of a list of fundamental rights, as described in the relevant European and International Charters, that are mostly impacted by the use of AI systems in the judicial sector; and (ii) the identification of the components of an hypothetical AI system used in this sector – for example, the training data, the algorithm, and the interface.

Following said assessments, the research team will single out a series of parameters and attributes that influence the impact of the AI systems used in the judicial field on fundamental rights and will translate them into machine-readable keywords. For example, one attribute that is very likely to have a significant impact on the fundamental right to equality and, in particular, non -

* The authors of this paragraph are Dr. Roberto Trasarti, Dr. Roberta Savella and Dr. Francesca Pratesi, with the contribution of Dr. Maria Cristina Gaeta.

discrimination, before the law is ethnicity (very often referred to as 'race' in the literature and legal documents), as we have seen in the debate surrounding the use of the software 'COMPAS'[17]. Therefore, there will likely be the need to include 'ethnicity' as a keyword for the prototype to analyse in its impact assessment on equality and non-discrimination. Moreover, it is important to also consider the proxy parameters that can be associated with the attributes identified. For example, numerous studies have concluded that the zip code is very often a proxy for 'ethnicity'[18].

Therefore, the first step of project will be the identification of the fundamental rights that are primarily relevant when applying AI systems in the justice sector, in order to ensure their protection. We considered three sources as starting point: the charter of Fundamental Rights of the European Union, the Universal Declaration of Human Rights and the European Convention on Human Rights, obtaining the following preliminary selection of fundamental rights that can be related to the judicial field.

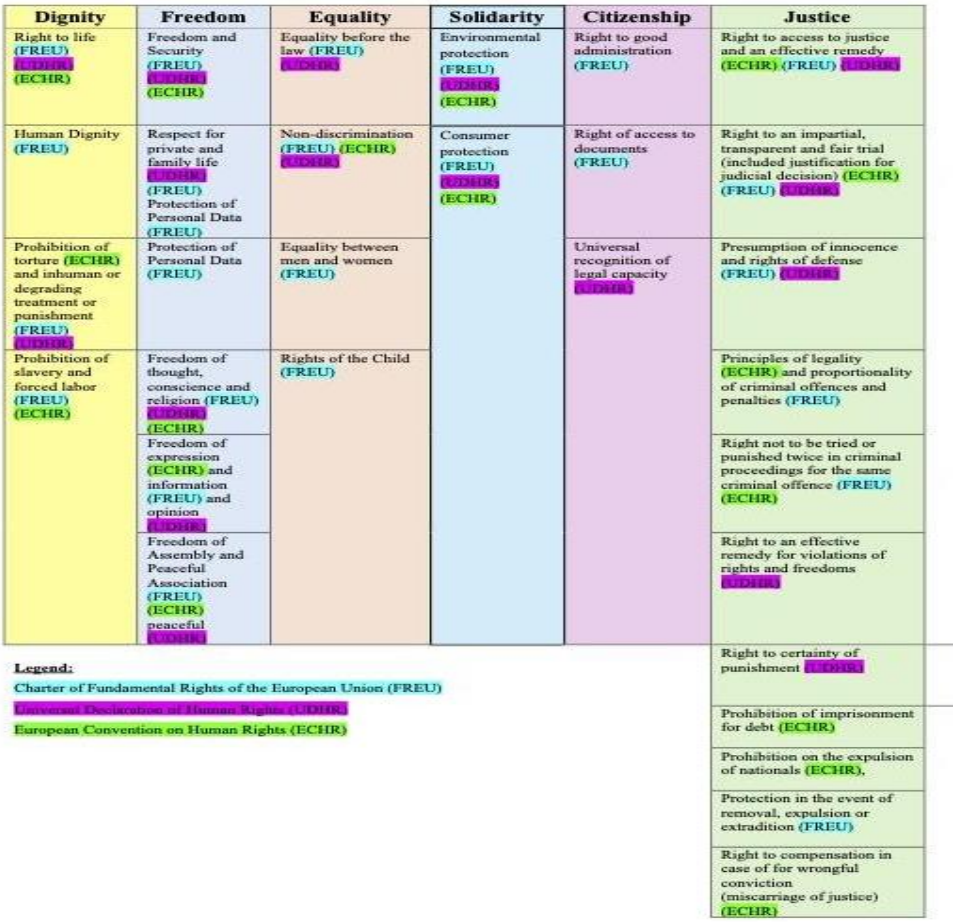

Image no. 2 – First selection of fundamental rights in the judicial filed

[17] J Angwin, J Larson, S Mattu, L Kirchner, 'Machine bias' (2016, 23 May) ProPublica' https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; A Lee, 'Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing' (2019, 19 February) UCLA Law Review, https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/.

[18] M Favaretto, E De Clercq, BS Elger, 'Big Data and discrimination: perils, promises and solutions. A systematic review' (2019) 12(6) J Big Data https://doi.org/10.1186/s40537-019-0177-4.

From this first set we want to identify specific protected attributes by examining the relevant laws (as shown in the example we provide in the next section) and therefore specific algorithms to extract quantifiable metrics which are relevant to evaluate the impact of the AI system on fundamental rights.

The second step of the experiment will be the planning and implementation of an empirical legal study (ELS) based on the impact of AI applications in the judicial field in concrete. It will then be necessary to play out concrete scenarios of using an AI system to test its operation. In this light testing environments of high-risk AI systems are of great importance, also in light of the AI Act provisions. More precisely, testing in real-world conditions outside AI regulatory sandboxes (art. 60-61 AI Act), as well as, more in general, testing activities in compliance with art. 9 AI Act, can guarantee simulations of operational environments in which to test the functioning of selected domains of high risk AI systems to complete the assessment. Indeed, experimentation through testing in real-world environments and sendboxes allows a way out of self-assessment, thus avoiding the problems in which all current evaluation systems (e.g., ALTAI or HUDERIA) are locked. In the testing phase, indeed, there will be the need to analyse each component of the theoretical AI system: the data used for training the algorithms, the model extracted and the output obtained by the application of the model on new data, the functioning of the AI system.

The third step of development will be the elaboration of metrics for each parameter for the concrete measurement of AI risks for the individuals. THe metrics will also be created through the individuation of algorithms to extract these metrics for assessing the impact of AI on the selected fundamental rights.

Finally, the fourth step will be the concrete and automatic measurement of the impact of AI on the selected parameters, using the individuated metrics (automatic evaluation of the results of the experiment). The team will then design a translation matrix that will use the outputs of the abovementioned steps, translating them into specific keywords to automatically assess the impact of the individual AI system on the fundamental rights previously selected. This operation will provide quantifiable metrics that will be used to calculate the overall impact on the fundamental rights of the AI system analysed.

7. Experiment design.

In order to show the approach, we present how we are addressing the problem in the case of equality and, more specifically, focusing on the right to non-discrimination.

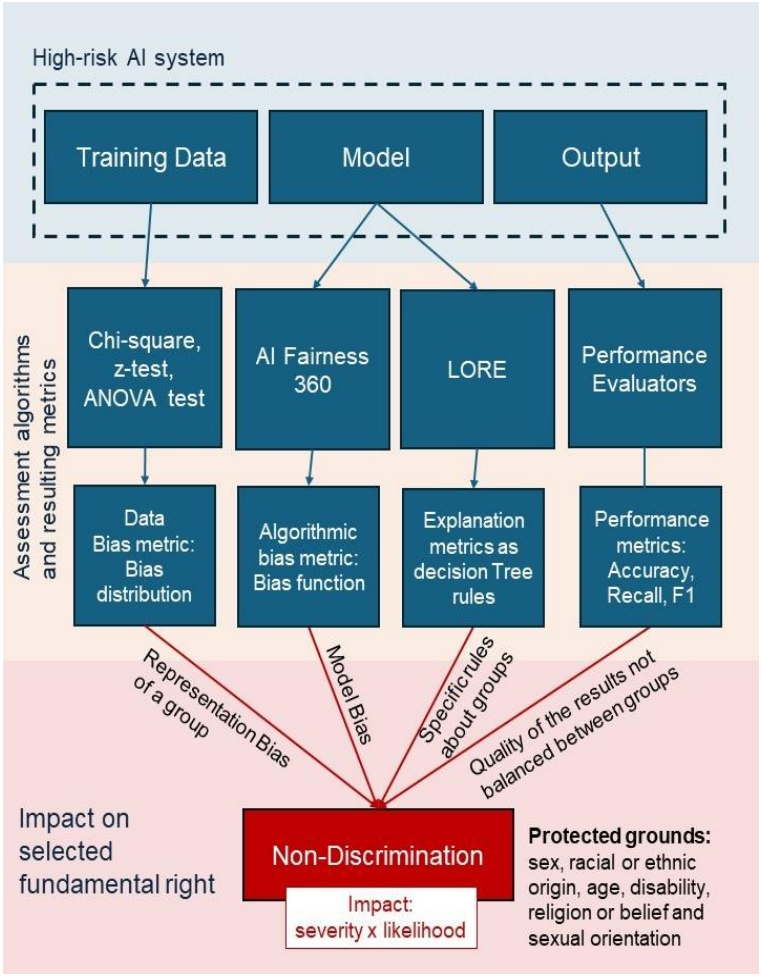

Image no. 3 –Assessment algorithms and resulting metrics for measuring the impact of high-risk AI on non-discrimination right.

We selected this right as literature has observed the discriminatory effects of biased systems[19] and it is particularly relevant for the use of AI in the judicial field, as shown by the COMPAS case[20]. Moreover, we believe that the selection of keywords in this context can be facilitated by the fact that European non-discrimination law prohibits differential treatment on the basis of a fixed and limited list of "protected grounds" i.e. "*characteristics of an individual that should not be considered relevant to the differential treatment or enjoyment of a particular benefit*"[21]. The protected grounds addressed by the European non-discrimination Directives are: sex, racial or ethnic origin, age, disability, religion or belief, sexual orientation. To design our prototype, we can use these protected grounds as keywords and for each keyword we will identify

[19] J Buolamwini, T Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018.

[20] J Angwin, J Larson, S Mattu, L Kirchner, 'Machine bias' (2016, 23 May) ProPublica' https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[21] European Union Agency For Fundamental Rights, "Handbook on European non-discrimination law – 2018 edition", February 2018, https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition.

quantifiable metrics to be measured with automated tools integrated in our prototype. For the purpose of this example, we selected the 'ethnicity' attribute, as it was particularly relevant in the COMPAS case[22]. We theorised a system composed by training data, model and output. To examine this system, we would need to apply algorithms for three purposes: bias detection and fairness auditing, explainability, and evaluation of the performance. To analyse the dataset, we could integrate in our model some statistical methods used in this field, like Chi-square, Z-test, and ANOVA test, to assess the bias distribution in the training data[23]. For bias detection in the model, we identified the open-source toolkit AI Fairness 360 [24] that could be integrated into our model through its publicly available APIs. This tool would identify representation biases in the system, allowing our prototype to calculate the impact on the right to non-discrimination. To obtain the explainability of the system we could apply LORE[25] or other existing explainable AI algorithms[26]. LORE is an agnostic method used to provide interpretable and faithful explanations for black box systems, and we could use it to find out the specific rules applied to the different ethnic groups. Specific metrics we would analyse for explainability are fidelity, faithfulness, completeness, comprehensibility, succinctness, and stability. Finally, we would need to apply a statistical evaluator to determine the accuracy, recall and F1 of the output, so we could evaluate whether the quality of the result is balanced or not balanced among the different ethnic groups.

8. Conclusions.

Both at the international and European levels, there is a clear understanding of the need to protect human beings from specific processes and products endowed with high-risk AI systems, and the need for such protection to be carried out to prevent damages, thus generating an AI assessment with regard to the fundamental rights as outlined in the relevant 'constitutional charters' of the EU and international framework. What is lacking is an effective risk assessment tool for deployers of high-risk AI systems. In this light, the study proposes a concrete FRIA implementation model, in line with the AI Act. Indeed, FRIA is envisaged in the European AI Act but, not specified in its structure nor in its implementation methodologies, and this is because the FRIA approach generates, from a legal point of view, conceptual and normative issues that affect the current system of the protection of human rights by autonomous agents.

---

[22] J Angwin, J Larson, S Mattu, L Kirchner, 'Machine bias' (2016, 23 May) ProPublica' https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[23] ER Girden (1992). ANOVA: Repeated measures. Sage Publications, Inc.

[24] N Mehrabi, F Morstatter, N Saxena, K Lerman, A Galstyan, "A Survey on Bias and Fairness in Machine Learning", arXiv:1908.09635v3 [cs.LG]

[25] R Guidotti, A Monreale, S Ruggieri, D Pedreschi, F Turini, F Giannotti, "Local Rule-Based Explanations of Black Box Decision Systems", arXiv:1805.10820v1 [cs.AI]

[26] D Rudresh, D Devam, N Het, S Smiti, O Rana, P Pankesh, Q Bin, W Zhenyu, S Tejal, M Graham, R Rajiv 'Explainable AI (XAI): Core Ideas, Techniques, and Solutions' (2023) 55(9) ACM Comput. Surv, Article 194 https://doi.org/10.1145/3561048.