European Review of Digital Administration & Law - Erdal
2022, Volume 3, Issue 2, pp. 117-131
ISSN 2724-5969 – ISBN 979-12-218-0798-1 – DOI 10.53136/9791221807981 09

*AI Systems in the Public Sector: Risks and Its Answers Within the EU Data Protection Framework**

Genoveva Gil García

(Tech-lawyer and data protection specialist working at Serendipity Holding B.V (Eindhoven, The Netherlands) and former Blue Book trainee at the Data Protection Office of the European Commission. She holds an LL.M in Law and Technology from Tilburg University, The Netherlands)

ABSTRACT In this paper the author analyses the risks posed by AI systems and the solutions already offered by the existing Data Protection Framework in the EU. In this regard, algorithmic risk-assessment tools are taken as case studies throughout the contribution. The analysis, although focused on Data Protection law, addresses the proposal for an AI act and takes into consideration the technicalities of AI systems. The paper concludes with some recommendations to be considered when implementing this new technology in our society, and especially in the public sector.

1. *Introduction*

The use of artificial intelligence is becoming more and more widespread in different sectors of society, including the public sector. For instance, there are already some discussions regarding the possibility of AI replacing judges or lawyers ('robot judges' or 'robot lawyers'), the use of AI as an assistance in the practice of law, the use of AI in job-recruitment processes[1] or even in the health sector.[2] These are just a few examples where the family of decision-making AI systems can be found.

Decision-making AI systems are designed to help in decision-making processes by mainly using automated data processing and machine-learning techniques. These self-learning AI systems make predictions or reach decisions by analysing large amounts of data and identifying patterns within datasets.[3] In particular, a great part of these tools conducts risk profiling by ranking individuals or groups, using correlations and probabilities drawn from the analysis of Big Data, to determine the level of risk of a certain event to occur.[4] Some examples can also be seen in the fintech sector, where the tools profile users into risk categories before providing financial advice;[5] in the insurance sector, where some models for assessing the risks of insurance companies' functioning have been explored;[6] in the context of criminal proceedings, to aid judges in the decision-making process, but also to grant prison privileges;[7] for unemployed profiling at public administration level;[8] or in the detection of tax fraud, where some of these tools have already been tested by some jurisdictions in the EU (e.g., Poland and the Netherlands).[9]

* Article submitted to double-blind peer review.

[1] For instance, *Pure Matching* is a recruitment matching AI system created by a software company in the Netherlands which promises to match available vacancies and jobseekers. See www.purematching.com/how-it-works accessed 17 April 2023.

[2] See C. Habib et al., *Health Risk Assessment and Decision-Making for Patient Monitoring and Decision-support using Wireless Body Sensor Networks*, in *Information Fusion*, vol. 47, 2018, 10-22.

[3] Committee of experts on internet intermediaries of the Council of Europe, Algorithms and human rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, 2018, 6-7.

[4] See S. van Schendel, *Risk Profiling by Law Enforcement Agencies in the Big Data Era: Is There a Need for Transparency?*, in E. Kosta et al. (eds.), *Privacy and Identity Management: Fairness, Accountability and Transparency in the Age of Big Data*, Springer, 2018, 275-289.

[5] See S. Krishnan, S. Deo and N. Sontakke, *Operationalizing algorithmic explainability in the context of risk profiling done by robo financial advisory apps*, in *Data Governance Network*, 2020.

[6] See O. Kozmenko and V. Oliyniyk, *Statistical model of risk assessment of insurance company's functionaning*, in *Investment Management and Financial Innovations*, vol. 12, 2015, 189-194.

[7] For instance, RisCanvi is currently being used in Catalonia (Spain) to estimate the risk that inmates reoffend when deciding whether or not to allow for parole. However, the system was not subject to any impact assessment and there is little transparency about it. (N. Bellio López-Molina, *In Catalonia, the RisCanvi algorithm helps decide whether inmates are paroled*, 2021).

[8] Joint Research Centre (European Commission), AI Watch. Artificial Intelligence in public services. Overview of the use and impact of AI in public services in the EU, Brussels, 2020, 46.

[9] See M. Papis-Almansa, *The Polish Clearing House System: A 'Stir'ing Example of the Use of New Technologies in Ensuring VAT Compliance in Poland and*

Indeed, in 2014, the Dutch government launched SyRi (*Systeem Risico Indicatie*), a tool which aimed to detect different forms of fraud, including social benefits, allowances, and tax fraud. The Dutch Tax Authority penalised families over a suspicion of fraud based on the risk scores provided by this AI risk-assessment tool. Many households – usually belonging to ethnic minorities or families with lower incomes– fell into poverty given the high amount of the fines imposed by the authority due to a wrong risk indicator. To date, a great part of them is still suffering from the economic consequences.[10] However, in February 2020, the Hague District Court ruled that the legislation regulating the use of SyRi, violates Article 8 of the European Convention on Human Rights (hereafter ECHR). The Court stressed that the application of SyRi was "*insufficiently transparent and verifiable*", and the authorities ceased to use this tool.[11]

This case illustrates that this new technology poses challenges to society and hence the need to address them. Therefore, it is necessary to define how to shape these tools in order to reap the benefits while protecting individuals rights and freedoms. In this regard, there is a need for an in-depth analysis of this topic from the perspective of the right to privacy and data protection since the processing of personal data is inherent to the nature of this technology.

Consequently, in this contribution, the author will try to briefly identify the risks posed by this new technology to individuals' fundamental rights and hence the challenges that lie ahead for (but not exclusively) public administrations. From there, it will be possible to explain how the instruments offered by the existing data-protection framework can help mitigate some of the risks presented by this technology, and how this can be translated into some considerations or recommendations for the implementation and use of AI systems by public administrations. The author will illustrate the latter by referring to some examples found in the public sector.

In the current legislative context, this analysis becomes even more important, due to the Proposal for a Regulation of Artificial Intelligence (hereafter AI act) of the European Commission of the 21 April 2021,[12] the recent Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence published on the 28 September 2022,[13] or the opening of negotiations for a Council of Europe Convention on Artificial Intelligence.[14] The proposal for an AI act establishes the requirements that providers and users of AI systems will need to fulfil in order to place them in the market or to put them into service. It follows a risk-based approach defining clear requirements for high-risk AI systems. Indeed, a tool like SyRi would be considered a high-risk AI system according to Article 5 of Annex III of the Proposal.[15] It remains to be seen what the final text of the Proposal will be, but the act is being widely debated not only by civil society organisations, tech lawyers and scholarship, but also in the European Parliament and the Council in accordance with the ordinary legislative procedure.

With regard to the methodology, the author will approach the topic from an EU perspective. For this, legal and non-legal scholarship will be useful to gain some insight into the topic as well as to identify potential risks and possible solutions to mitigate them. The study of already-existing examples of risk-assessment tools (e.g., SyRi), will be taken into consideration. This will help the author to get familiar with this technology and to identify the risks and answers to the challenges.

The use of European policy documents will

Selected Legal Challenges, in *EC Tax Review*, vol. 28, 2019, 43-56; and S. van Schendel, *The challenges of Risk Profiling Used by Law Enforcement: Examining the Cases of COMPAS and SyRi, in Regulating New Technologies*, in L. Reins (ed.), *Uncertain Times*, The Hague, 225-240.

[10] See *Dutch scandal serves as a warning for Europe over risks of using algorithms* www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/ accessed 24 October 2022.

[11] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[12] European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, Brussels, 2021.

[13] European Commission, *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*, Brussels, 2022.

[14] European Commission, *Recommendation for a Council Decision authorising the opening of negotiations on behalf of the European Union for a Council of Europe convention on artificial intelligence, human rights, democracy, and the rule of law*, Brussels, 2022.

[15] European Commission, *Annexes to the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, Brussels, 2021.

be required to describe the current views of different European institutions on the topic. In this regard, not only will the papers issued by the European Union be looked into, but also the ones from the Council of Europe. Indeed, in order to propose possible solutions, the Ethics Guidelines for Trustworthy AI will be a helpful starting point to rely on, as it sets out the principles that the use of AI should respect.[16] As indicated, European data-protection law will also be of utmost relevance due to the vast amount of personal data usually processed by these tools. Furthermore, the proposal for an AI act will be briefly addressed since it contains the requirements that high-risk AI systems will have to comply with.

2. *AI systems, machine learning and risks of algorithmic risk-assessment tools*

2.1. *Introduction*

There is no common definition for "artificial intelligence", "algorithms" or "AI systems". However, the European Union Agency for Fundamental Rights (hereafter FRA) defines 'algorithms' as "a sequence of commands for a computer to transform an input into an output".[17] To put it simple, algorithms are part of so-called AI systems, which the Commission in the Proposal for an AI act has defined as: "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". This definition, among quite a few other aspects, is currently being debated in the Parliament and the Council.

A great majority of AI systems make use of machine-learning techniques, which means that they 'learn' by analysing large amounts of data in order to establish correlations within datasets. Hence, machine learning is a necessary component of AI system models and AI.

As indicated in the Introduction of this paper, a great part of decision-making AI systems makes use of these techniques to conduct risk profiling and to reach a decision according to the risk score concluded by the system. This specific type of AI systems is the subject of this contribution, and it is referred to as algorithmic risk assessment tools.

Machine-learning AI systems require at least three different types of datasets to produce a certain outcome: training data, input data and inferred labels.[18] Training data are used to build the model. Input data are the information introduced into the AI system to achieve the desired output. Finally, the system finds correlations between the training and input data, and it produces an inferred label. In the case of SyRi, due to its lack of transparency it was not clear whether it made use of machine-learning techniques or not. Actually, the Court of The Hague noted that the State did not disclose the risk model and the indicators composing the tool .[19] However, if the tool would have followed the latter structure, training data would be constituted by historical data of former fraudsters; input data would be potential fraudsters' personal data; and the output would be a risk score based on a correlation between the two data sets.

This process shows the great influence and importance that data quality has in risk profiling conducted by the tool. If training data are of low quality or contain biases, it may lead to inaccurate outputs which could infringe fundamental rights, like the right to non-discrimination or the right to privacy and data protection.[20]

On another note, machine-learning AI systems are commonly known as 'black boxes'. However, this claim does not respond to all different models in which AI systems can be presented. For the purposes of this paper, machine-learning AI systems can be divided into interpretable models and deep-learning models. Interpretable models (e.g., decision trees) provide transparency and allow human users to trace the steps taken by the tool in the decision-making process. Deep-learning models (e.g., neural networks) are considered 'black boxes' either because their complicated structure and functioning are

[16] High-Level Expert Group on Artificial Intelligence of the European Commission, *Ethics Guidelines for Trustworthy AI*, 2019.

[17] European Union Agency for Fundamental Rights, *#BigData: Discrimination in data-supported decision making*, Vienna, 2018, 4.

[18] European Union Agency for Fundamental Rights, *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*, Vienna, 2019, 4-5.

[19] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 paragraph 6.49 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[20] European Union Agency for Fundamental Rights, *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*, 5.

uninterpretable to human users, or because only a few experts are capable to understand the machine codes.[21] In addition, sometimes the opacity is not only due to technical reasons, but because the tool contains proprietary know-how and hence it is protected by intellectual property rights. This is why it is widely said that 'black-box' algorithms cause a lack of transparency or explainability in the decision-making process.[22] There is a clear risk here for the right to data protection. According to Data-protection legislation, the data subject has the right to know what personal data and the way in which data are processed. Equally, in the case of SyRi, individuals should have also been provided with this information in order to be able to challenge the decision adopted by the tool.

In this same regard, in 2012 the Polish Ministry of Labour and Social Policy implemented an automated profiling system for unemployment. This system divided unemployed persons in three categories to determine the type of program they were eligible for. However, in this case citizens were informed neither of the score received, nor of how the tool reached this result. The tool was ruled unconstitutional by the Polish Constitutional Tribunal in 2019.[23]

In addition to the above, algorithmic risk-assessment tools make use of automated decision-making systems (hereafter ADMSs). ADMSs can be used to produce outcomes without human intervention, meaning that the decision would be fully automated; or to serve as a tool for humans in their decision-making process. In principle, the results provided by these tools should be considered as a mere instrument to aid the reviewer of a case during an investigation or procedure. However, the automation element will still be present and there is a risk of human reviewers being partial.

To conclude this section, by describing the tool it is possible to briefly identify the main risks that the use of AI risk assessment tools entails. These risks are essentially: risk of discrimination, risk of AI systems' opacity and the risk of falling into automation. The following sections will provide a more extensive description of them.

2.2. *The risks of algorithmic risk-assessment tools*

2.2.1. *Risk of discrimination*

Algorithmic risk-assessment tools are machines controlled and designed by humans. Consequently, the choices about data made by their designers will necessarily have an impact on the tool's prediction. These tools are trained on historical data and hence there is a risk of perpetuating and reinforcing historical biases or prejudices, which could lead to discriminatory outcomes.[24] This result would be contrary to the principle of fairness stated by the High-Level Expert Group on AI. According to this group, an AI system is fair if it is free from bias, discrimination, and stigmatisation.[25]

It is worth mentioning the distinction between direct and indirect discrimination provided by the Council of Europe in the context of ADMSs. Direct discrimination occurs when the decision about an individual is directly based on protected grounds such as race, ethnicity, or gender. Since these unfair biases are usually made sub-consciously, it is said that AI systems can exclude those biases. Indirect discrimination arises when a certain factor occurs more frequently among the groups against whom it is unlawful to discriminate. In this case, certain individuals are treated differently because the decision relies on biased data.[26]

For instance, this is the kind of discrimination which has been claimed in the case of COMPAS, a tool used by US courts to assess defendants' risk of recidivism when judges must determine the sentence for an individual.[27] This tool raised concerns about

[21] A. Rai, *Explainable AI: from black box to glass box*, in *Journal of the Academy of Marketing Science*, vol. 48, 2020, 138.

[22] Council of Bars and Law Societies of Europe, *Considerations on the legal aspects of artificial intelligence*, CCBE, 2020, 12.

[23] Joint Research Centre (European Commission), *AI Watch. Artificial Intelligence in public services. Overview of the use and impact of AI in public services in the EU*, Brussels, 2020, 46-47.

[24] L. Edwards and M. Veale, *Enslaving the Algorithm: From a "Right to and Explanation" to a "Right to Better Decisions"?*, in *IEEE Security & Privacy*, vol. 16, 2018, 46.

[25] High-Level Expert Group on Artificial Intelligence of the European Commission, *Ethics Guidelines for Trustworthy AI*, 2019, 12.

[26] Committee of experts on internet intermediaries of the Council of Europe, *Algorithms and human rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, 2018, 26-27.

[27] See Partnership on AI, Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System, 2019; Thomas Blomberg et al., *Validation of the COMPAS risk assessment classification instrument*, in *Centre for Criminology and Public Policy Research*

its fairness for being based on factors that seem biased, such as racial bias or gender bias.[28] In the case of SyRi, having dual nationality was connected to "low income" and interpreted as a risk indicator.[29] Similarly, the Polish unemployment tool assessed women in a different way than men, and the distinction lowered their chances to receive assistance from public authorities.[30]

The challenges to correctly design risk-assessment tools can be divided in two. First, the selection of the data that will be embedded into the machine. Second, the detection and avoidance of possible miscodes or errors that the AI system may fall into during the decision-making process.[31]

With regard to the first challenge, the quality of the data to feed the AI system is crucial to avoid the risk of discrimination. In this vein, the FRA highlights two sources of error when selecting the data: measurement errors and representation errors. Measurement error refers to "how accurately the data used indicate or reflect what is intended to be measured". For instance, if there is an intention to measure the country of origin of individuals, and this information is not available, their nationality could be used as a proxy. However, this proxy does not seem to be accurate enough to determine the country of origin of certain individuals.[32] Representation error concerns the question of how representative the sample population is. If some groups of the general population are not sufficiently represented in the sample, the output could be incorrect and biased. The FRA also highlights the importance of timeliness of the training data; in other words, training data should represent individuals at the present time.[33]

In this respect, the question arises regarding which type of data should be included in the machine to avoid bias and hence discrimination (de-biasing data). The issue is that detecting and eluding discrimination is not straightforward. Indeed, it has been concluded that AI systems designed to be neutral can still produce discriminatory outcomes; that is to say, the risk will not be easily solved solely by removing the information directly referred to protected grounds (e.g., race, ethnicity, or gender). In fact, there might be some proxies or residual information which still refer to individuals' protected attributes.[34]

Završnik raises an interesting debate about whether it is desirable or not to conduct de-biasing procedures. He considers that this would not be suitable if choices about data are taken "behind closed doors by computer scientists in a laboratory".[35] Then he poses the question about whether our society would prefer human bias or machine bias. Although the latter discussion would be out of the scope of this paper, it is worth mentioning that the design of the AI system should involve not only computer experts, but also lawyers; and, in any case, the process should be transparent and not happen "behind closed doors". In the author's view, human overview and transparency throughout the whole process are key elements for preventing discrimination and biased outcomes. In this regard, the FRA highlighted the relevance of periodically auditing AI systems. Although there still is little research on which datasets provoke discriminatory predictions, there are already methods to detect which information contributes most to AI systems' outcomes.[36] This is the first step to improve AI systems' fairness.

As with the presence of biased data, AI systems' technical miscodes or errors can also be responsible for discriminatory outcomes. These errors should also be solved in the design process, since they could increase the unequal rates of 'false positives' and 'false negatives'. However, it should be remembered that this is not always due to wrong codes

(Florida State University), 2010.

[28] See M. Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, in *American Criminal Law Review*, vol. 56, 2019, 1553-1577; M. Hamilton, *The sexist algorithm*, in *Behavioral Sciences & the Law*, vol. 37, 2019, 145-147.

[29] See "*Dutch scandal serves as a warning for Europe over risks of using algorithms"* www.politico.eu /article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms accessed 17 April 2023.

[30] "*Poland: Government to scrap controversial unemployment scoring system"* https://algorithmwatch.org /en/poland-government-to-scrap-controversial-unemployment-scoring-system accessed 17 April 2023.

[31] In a similar vein, see A. Završnik, *Algorithmic justice: Algorithms and big data in criminal justice settings*, in *European Journal of Criminology*, vol. 18, 2019, 623-642.

[32] European Union Agency for Fundamental Rights, *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*, 11.

[33] *Ibid*., 12.

[34] European Union Agency for Fundamental Rights, *#BigData: Discrimination in data-supported decision making,* 8.

[35] A. Završnik, *Algorithmic justice: Algorithms and big data in criminal justice settings*, 633.

[36] European Union Agency for Fundamental Rights, *#BigData: Discrimination in data-supported decision making*, 6.

programmed in the machine, since the presence of biased factors is still one of the main reasons of these disparate results.[37]

2.2.2. *Risk of AI systems' opacity*

As previously stated, it is possible to differentiate between explainable and interpretable AI models (e.g., decision trees) and deep-learning 'black-box' AI models (e.g., neural networks). Both pose a risk of discrimination, but 'black-box' models also carry a risk of opacity and a lack of explainability. They are known as 'black boxes' because the decision-making process remains opaque. This opacity may be due to three different reasons which were mentioned before, and which Burrell names as: intentional opacity, illiterate opacity, and intrinsic opacity.[38] Intentional opacity refers to algorithms that are protected by trade secrets or intellectual property rights (e.g., COMPAS).[39] Illiterate opacity arises when the tool is only understandable for computer scientists who can read machine codes. Finally, intrinsic opacity refers to AI systems which are uninterpretable to any human user. All combinations between these types of opacity are possible.

In all three cases, the lack of transparency and explainability jeopardises citizens´ right to data protection. For instance, individuals may not be able to challenge an administrative resolution if part of its reasoning is opaque (e.g., Polish unemployment profiling tool). Additionally, intrinsic opacity constitutes a barrier for designers and developers of the AI system. This is perhaps the most worrying form of opacity, since no person would be able to explain how a certain prediction was made. Hence, it will not be possible to detect potentially biased outcomes, nor will it be possible to initiate a de-biasing procedure.

Having said this, it is necessary to explain the difference between transparency and explainability in the context of AI systems. In this regard, a study conducted by the European Parliament Research Services is very enlightening.[40] According to it, transparency is the availability of the AI model's code, design documentation and learning dataset. However, it does not mean that it is available to the public. As for explainability, it is the availability of explanations about the logic behind the AI systems' decision.

Bearing in mind the latter, it seems that only a transparent AI system can properly address the risks posed to citizens' fundamental rights. Indeed, transparent, and explainable AI systems should be the final aim. For citizens' rights not to be hampered, there is a need for justifications to the reasoning made by the tool. A mathematical or technical explanation of how the algorithm evolved from the input to the output will not suffice.[41]

Therefore, the question is how to implement 'explainability'. The aforementioned European Parliament's study describes three possible approaches:[42] a black-box approach, a white-box approach and a constructive approach. The first one analyses the relationship between the inputs and outputs of the AI system without any knowledge of its code. The white-box approach considers that analysing the code is feasible. Lastly, the constructive approach operates by inserting explainability requirements in the design process of the tool.[43]

In this vein, there are already some techniques which address these approaches. For instance, so-called model-specific techniques, which incorporate interpretability within the structure of 'black-box' models; or model-agnostic techniques which use the inputs and predictions of the 'black box' to produce explanations (explainable AI (XAI)).[44]

[37] To see a graphic example on disparate results between two groups due to the presence of biased factors, see R. Courtland, *The bias detectives*, in *Nature*, vol. 558, 2018, 357-360.

[38] J. Burrel, *How the machine "thinks": Understanding opacity in machine learning algorithms*, in *Big Data & Society*, 2015, 2.

[39] COMPAS originated the *Loomis v. Wisconsin* case, in which Loomis argued that using predictive algorithms violated his right to due process because they did not allow him to verify the scientific validity and accuracy of such algorithms. (See Taylor R. Moore, *Trade Secrets Algorithms as Barriers to Social Justice* in *Center for Democracy and Technology*, and Council of Bars and Law Societies of Europe, *Considerations on the legal aspects of artificial intelligence* (CCBE), 2020, 24.

[40] European Parliament Research Services (Panel for the Future of Science and Technology), *Understanding algorithmic decision-making: Opportunities and challenges*, 2019, 3.

[41] For the distinction between justification and explanation, see S. Quattrocolo, *An introduction to AI and criminal justice in Europe*, in *Revista Brasileira de Direito Processual Penal*, vol. 4, 2019, 1528.

[42] These approaches will be further elaborated in section 3.2 "transparency as a means to explainability".

[43] European Parliament Research Services (Panel for the Future of Science and Technology), *Understanding algorithmic decision-making: Opportunities and challenges*, 2019, 4.

[44] A. Rai, *Explainable AI: from black box to glass box*,

From the above, it can be concluded that the main risk with regard to the tool's opacity is the lack of explainability. In fact, although full transparency of the code and mechanism of the AI system is desirable, in order to exercise their rights, citizens may also need an explanation of the logic involved behind the AI system. However, in any case, it is quintessential that the system also be transparent. Indeed, transparency should be seen as a means to an end.[45] The end is having an explainable and unbiased AI system which respects fundamental rights like the right to privacy and data protection. Consequently, it is required that the entire algorithmic process be transparent so as to enable regulators, designers, auditors, deployers and developers, to detect and address its flaws. This is the first step to ensure the implementation of a fair AI system.

2.2.3 *Risk of falling into automation*

Algorithmic risk-assessment tools are part of ADMSs. Therefore, taking SyRi´s example, there will be an automation component in the decision-making process which may jeopardise decision makers' discretion. If this occurs, the right to privacy and data protection will be affected since the risk score issued by the tool can interfere in citizens´ private life. In this regard, these tools should be conceived as an additional element to aid decision makers in reaching a decision. In fact, the opposite would be in breach of Article 22 of the GDPR (right not to be subject to a decision based solely on automated processing, including profiling).

The European Commission strongly affirms that AI should follow a human-centric design approach.[46] This is especially important in the context of public services, where citizens should be put at the centre. As it was mentioned before, human overview throughout the whole process is a key element to address the different risks posed by this kind of tools. The "surveillance" during the design process by IT experts and data protection experts would act as a safeguard for citizens' fundamental rights. This would be a crucial element to detect possible biases or errors produced by the tool.

However, the question here is how to ensure that decision makers do not fully rely on the decision given by the tool. Indeed, human overview cannot mean a decision maker "just signing off the recommendations or outputs from an algorithm".[47] Ultimately, the problem lies in the so-called 'control problem', which states that humans tend "to fall into automation complacency and bias once the system operates reliably most of the time".[48]

In the case of the Polish AI system, the technology was initially projected as an advisory tool for public servants as decision makers of a case. However, it turned out that decision makers overrode less than 1 in 100 decisions.[49]

3. *EU Data Protection Framework: instruments to address the risks of algorithmic risk-assessment tools*

3.1. *Introduction*

The right to privacy and the right to personal-data protection are enshrined in Articles 8 of the ECHR and Articles 7 and 8 of the EU Charter of Fundamental Rights. They are closely related to each other since they both strive to protect individuals' autonomy and human dignity. However, they differ in their scope and formulation. The right to privacy –referred to in Articles 7 of the EU Charter and 8 of the ECHR as the right to respect for private and family life– is invoked whenever an interference in the individual's private sphere has occurred. By contrast, the right to personal-data protection is broader since it comes into play whenever personal data are being processed.[50] In this regard, risk-assessment tools are likely to process personal data, e.g., SyRi processed large amounts of personal data which included, *inter alia*, work

138.

[45] This idea was expressed by J. Cobbe at the 14th International Conference Computers*, Privacy & Data Protection: Enforcing Rights in a Changing World* (27-29 January 2021).

[46] European Commission, *White Paper on Artificial Intelligence – A European approach to excellence and trust*, Brussels, 2020, 3; and European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, Brussels*, 2021, 1.

[47] *Mutatis mutandis*: European Union Agency for Fundamental Rights, *Getting the future right. Artificial Intelligence and Fundamental Rights*, Vienna, 2020, 64.

[48] J. Zerilli et al., *Algorithmic Decision-Making and the Control Problem*, in *Minds and Machines*, vol. 29, 2019, 565.

[49] Joint Research Centre (European Commission), *AI Watch. Artificial Intelligence in public services. Overview of the use and impact of AI in public services in the EU*, Brussels, 2020, 47.

[50] European Union Agency for Fundamental Rights, *Handbook on European data protection law*, Vienna, 2018, 19-20.

data, education data, personal identification data (e.g., name, address, city).[51] Consequently, this kind of AI tools would trigger the application of data protection rules and citizens would be considered as data subjects.

In this regard, the processing of personal data by these tools would require having a legal basis as per Article 6 of the GDPR. In the case of SyRi, there was a legal basis enshrined in Dutch law,[52] but this was certainly not sufficient to guarantee the right to privacy and data protection. In the following sections, different instruments offered by the existing Data Protection Framework will be analysed in light of the risks previously identified. This analysis will provide some guidance on how to address these challenges and on how to work towards the implementation of human-centric AI systems in the public sector.

3.2. *Transparency as a mean to explainability*

The principle of transparency is one of the core principles regulated by the GDPR in Article 5. This requirement is quintessential to the rights concerning the processing of people's personal data (e.g., right of access, right to rectification or erasure of personal data). Indeed, this requires the information shared with data subjects "to be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used" (Recital 58 of the GDPR).

Transparency addresses the risk of opacity, which hampers the right to data protection. In concrete, regarding SyRi, the Hague District Court, in February 2020, delivered a judgement where it decided that SyRi legislation did not comply with Article 8 of the ECHR mainly because of lack of transparency. The risk model and the risk indicators were 'secret' and the legislation provided no duty to inform data subjects that their data were being processed in SyRi.[53] Consequently, there is a need to make 'black-box' systems transparent. In this regard, transparency is a means to achieve the aim of explainable and unbiased AI systems pursued by citizens' fundamental rights.

There are already different methods and approaches to tackle the 'black-box' issue as already introduced in sub-section 2.2.2. The European Parliament Research Services distinguishes between three approaches to an explainable AI system. First, the black-box approach which demands to explain complex and difficult AI systems without any knowledge of the code, in other words: without "opening" the 'black box'. Second, the white-box approach makes it possible to analyse the code of the system by providing explanations to a wider range of AI systems (e.g., deep neural networks). Last, the constructive approach refers to situations where explainability is built during the design process of the AI system.[54]

Both the black-box and white-box approaches consist of explaining 'black-box' systems with separate explanation models. These models help to make those AI systems explainable. For instance, Local Interpretable Model-Agnostic Explanation (LIME) works by implementing an interpretable model to a specific outcome produced by an opaque system.[55] Conversely, the constructive approach aims to build an inherently interpretable model without requiring second models to explain the 'black box'. Thus, it consists of "originally" interpretable and transparent AI systems (see sub-section 2.2.2).

Having said that, the scientific community has drawn attention to the trade-off between explainability and accuracy. In this sense, a big part of the research community advocates for the reliance on the first two approaches. This scholarship considers that despite how easy it is to build intrinsically explainable models, the simpler these models are, the less accurate their results will be. Therefore, their proposal is to build complex but highly accurate 'black-box' models and then explain their inner functioning with second *post-hoc* models.[56] This entails that a certain degree of

[51] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 paragraph 4.17 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[52] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 paragraph 4.8 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[53] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[54] European Parliament Research Services (Panel for the Future of Science and Technology), *Understanding algorithmic decision-making: Opportunities and challenges,* 2019, 48-52.

[55] Information Commissioner's Office and The Alan Turing Institute, *Explaining decisions made with AI*, 2020, 124.

[56] See Sarkar et al., *Accuracy and interpretability trade-offs in machine learning applied to safer gambling*, in CEUR Workshop Proceedings, vol. 1773, 2016; Z.C.

transparency and explainability be provided once the model has already been deployed. On the other hand, another part of the literature supports the constructive approach where transparency is provided in the design process before the implementation of the model.[57]

Rudin is of the opinion that, although challenging, it is possible to design interpretable models which provide their own explanations while also being accurate. Indeed, she clarifies that using second *post-hoc* models entails the risk that "any explanation method for a black-box model can be an inaccurate representation of the original model in parts of the feature space". [58]

In light of the above, opting for a constructive approach would be the best option to meet the requirements of transparency and explainability. If risk assessment tools are based on an interpretable model, data subjects, would be in a better position to understand how the tool reached the risk score as well as which attributes it took into consideration, and hence they will be able to know how their personal data are being processed. This is the type of information that was not provided to citizens being subject to the Polish unemployment tool. Individuals' inability to understand how the tool had reached the score made it difficult for them to later challenge the administrative decision. Indeed, Rudin affirms that it is easier to detect and avoid possible bias and data privacy issues within an interpretable model than within a 'black box'.[59] For this reason, it is necessary to invest in research and design processes of AI systems so to implement an explainable while accurate interpretable model. In this regard, this would be a decision of "data protection by design" where transparency could be effectively embedded into the tool (see section 3.3).

Nevertheless, this explainability strategy does not solve the issue with intentional 'black-box' systems where there is a trade secret or intellectual property right over the technology. Indeed, the model could be interpretable but still protected by intellectual property rights. In this regard, Recital 63 of the GDPR mentions that the right to access personal data "should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property". As the Norwegian Data Protection Authority (hereafter Norwegian DPA) suggests, a balanced solution could be achieved by providing data subjects with the information they need to protect their interests, while not disclosing trade secrets.[60] Therefore, the trade-off between IP rights and the requirements of transparency and explainability is an interesting angle to address in future research.

Having addressed how transparency and explainability can be technically achieved, it is useful to assess what the explainability requirement should include. In this regard, Wachter et al. distinguish between two kinds of explanations to be provided when an automated tool is involved. The first one refers to the system functionality which includes the general functionality of the automated decision-making system, the models, the logic, or the classification structures. The second one refers to the specific decisions, which are the rationales, reasons and individual circumstances that led to a concrete automated decision. The latter kind of explanation is the one referred to in Recital 71 of the GDPR ("…to obtain an explanation of the decision reached after such assessment and to challenge the decision"). Then, these authors differentiate between an explanation given *ex ante* and *ex post* automated decisions.[61] It should be noted that although Article 15 of the GDPR on the right of access does not mention a specific timing for the exercise of this right, the explanation should be given *ex ante* and *ex post* to the processing of the data by the tool.

According to Articles 13 (information to be provided where personal data are obtained

Lipton, The Mythos of Model Interpretability in Machine Learning, the concept of interpretability is both important and slippery, in *Association for Computing Machinery*, 2018; B. Lepri et al., *Ethical machines: The human-centric use of artificial intelligence*, in *iScience*, vol. 24, 2021.

[57] See R. Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmision*, in *KDD '15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, 1721-1730; B. Letham et al., *Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model*, in *The Annals of Applied Statistics*, vol. 9, 2015, 1350-1371; C. Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, in *Nature Machine Intelligence*, vol. 1, 2019, 206-215.

[58] C. Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 207.

[59] *Ibid*., 208.

[60] Datatilsynet (The Norwegian Data Protection Authority), *Artificial intelligence and Privacy*, 2019, 19.

[61] S. Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in *International Data Privacy Law*, vol. 7, 2017, 78.

from the data subject), 14 (information to be provided where personal data have not been obtained from the data subject), and 15 (right of access) of the GDPR, the data subject has the right to be informed of the purposes and categories of personal data that are subject to the processing activity. The processing must be lawful, fair, and transparent, which entails that the data subjects should be aware of how their data are being processed. From this, it could be determined that the scope of the explanation should include both the system functionality and the specific decisions of the AI tool. In this vein, the Information Commissioner's Office (hereafter ICO) identifies different ways to explain AI decisions, which include a rational explanation, a data explanation, and a fairness explanation.

The rationale explanation refers to the reasons behind a certain decision so to allow individuals to challenge the decision in a proper form. These are the specific decisions of AI tools explained in an accessible and non-technical manner.[62] The data explanation should inform individuals not only about what data have been used and how, but also about what other types of data have been used to design, train and test the AI model.[63] This entails providing information on the quality of the data so to prove that data sets are free from bias and that the training data are periodically verified and tested. Therefore, this explanation encompasses the input data (personal data of the individual) and the training data (historical data). In this regard, the proposal for an AI act includes among the information to be provided to users: "specifications for the input data, or any other information in terms of the training, validation and testing data sets used".[64] Last but not least, the fairness explanation consists of fostering trust among individuals subject to automation by informing them about the steps taken to design and implement an AI model which is unbiased, fair, and non-discriminatory.[65]

In a sense, it can be said that the last two kinds of explanations defined by the ICO offer information on the system functionality. However, it should be mentioned that technical or mathematical explanations should be provided to the extent necessary to understand the logic of the tool and to determine how it finds correlations and patterns within the dataset. In fact, the Norwegian DPA opines that "it is not always necessary to provide a thorough explanation of the algorithm, or even include the algorithm".[66] Therefore, the author of this paper understands that the information on the system functionality should be provided in order to allow data subjects to readily understand how the decision was made without delving into unnecessary technical aspects. Consequently, such a transparent explanation would enable individuals to verify whether their data are being processed fairly and lawfully. If this is not the case, they could submit a data-protection request to the controller of their personal data or even lodge a data-protection complaint before the corresponding Data Protection Authority.

3.3. *Privacy and data protection by design or "AI systems by design"*

Article 25 of the GDPR establishes the obligation of the controller "to implement appropriate technical and organisational measures (…) which are designed to implement data protection principles". This entails that the controller (the public authority implementing the AI system) is required to have data protection designed into the processing of personal data. In this sense, unlinkability, transparency and control over the data constitute entry points and goals to privacy by design processes.[67]

Consequently, in the context of risk assessment tools, this provision mandates that transparency be implemented in the design process of AI technology. In this regard, as it has been analysed throughout this contribution, the choice of an explainable AI system should be made during the design process of the tool so to comply with this legal requirement. In other words, opting for a constructive approach where explainability is built during the design process of the AI system. Indeed, the AI act highlights the relevance of design choices of high-risk AI systems.[68] As Bryson indicates, the extent to which transparency is embedded into a

[62] Information Commissioner's Office and The Alan Turing Institute, *Explaining decisions made with AI*, 2020, 20 and 23.
[63] Ibid., 25-26.
[64] Article 13 Proposal for a Regulation of AI.
[65] Information Commissioner's Office and The Alan Turing Institute, *Explaining decisions made with AI,* 28-29.
[66] Datatilsynet (The Norwegian Data Protection Authority), *Artificial intelligence and Privacy*, 21.
[67] Spanish Data Protection Authority (AEPD), *A Guide to Privacy by Design*, 2019, 13 www.aepd.es/sites/default/files/2019-12/guia-privacidad-desde-diseno_en.pdf accessed 17 April 2023.
[68] Article 10 Proposal for a Regulation of AI.

product constitutes a design decision and it is perfectly possible to achieve that the technology is designed to comply with laws.[69]

Having said the above, the design process of algorithmic risk-assessment tools is important to provide the tool with transparency and explainability, but it is also crucial to avoid other risks that would threaten citizens' fundamental rights. On the one hand, an adequate privacy by design strategy will ensure that personal data are stored in a secure manner and that data-protection principles be respected. On the other hand, a robust and careful design will also address the risk of biased outcomes by selecting the correct quantity and quality of the training data, or by developing a machine code that avoids undesired results. In this regard, the selection of an inherently interpretable model could already enable regulators, designers, auditors, deployers and developers, to detect and address its flaws before its implementation. To this should be added the importance of documenting the design decisions, not only to be able to provide a full explanation of how the tool reached a certain decision, but also to demonstrate compliance of the AI system with the requirements set out in the Proposal.[70] Moreover, this Proposal requires that these tools include record-keeping or logging capabilities that enable traceability of their functioning.[71] The latter requirements will facilitate audits and monitoring of the technology to correct possible errors or flaws.

In the field of data protection, it is considered that the design must be kept "user-centric" to guarantee the rights and freedoms of the users whose data are processed.[72] Similarly, the European Commission strongly affirms that AI should follow a human-centric design approach,[73] which centres individuals' needs, motivations, emotions, or behaviour in the development of the design.[74] As a consequence, and especially in the public sector, every design decision of the AI tool should be inspired by this approach with the aim to preserve citizens' fundamental rights ('citizen-centric approach'). Therefore, the author considers that the decision to opt for an inherently explainable AI system places humans in the centre. In any case, it would be advisable to adopt a participatory design[75] where computer scientists, engineers and mathematicians work closely with bar associations, lawyers, data-protection experts, or civil-society organisations in the design of the prospective algorithmic risk-assessment tool. Moreover, citizens' perspective should also be taken into consideration. This can be achieved through the establishment of an AI register (see also 3.5), conducting public campaigns informing about the initiation of an AI system project, collecting feedback from citizens on the prospective objectives and design of the tool… In this way end-users would be involved in the design process of the technology and their interests could be reflected in the final architecture of the tool. This will also help public authorities decide on whether it is feasible or not to proceed with the different phases of a concrete project.

These design options and decisions can be tested through regulatory sandboxes. The latter constitutes an interesting mechanism introduced by the Proposal for an AI act where public authorities as regulators together with innovators can test AI systems in a safe environment before placing them on the market or putting them into service. The first regulatory sandbox on AI was presented on June 2022 by the government of Spain and the European Commission. National authorities from other Member States should also be encouraged to do so.[76]

[69] J. Bryson, *The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation*, in M.D. Dubber et al (ed.), *The Oxford Handbook of Ethics of AI*, in Oxford Handbooks Online, 2020, 5.

[70] Article 11 Proposal for a Regulation of AI.

[71] Article 12 Proposal for a Regulation of AI.

[72] Spanish Data Protection Authority (AEPD), *A Guide to Privacy by Design*, 2019, 10 www.aepd.es/sites/default/files/2019-12/guia-privacidad-desde-diseno_en.pdf accessed 17 April 2023.

[73] European Commission, *White Paper on Artificial Intelligence – A European approach to excellence and trust*, 2020, 3; and, European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, Brussels, 2021, 1.

[74] See B. Shneiderman, *Human-Centered Artificial Intelligence: Three Fresh Ideas*, in *AIS Transactions on Human-Computer Interaction*, 2020.

[75] J. Auernhammer, *Human-centered AI: The role of Human-centered Design Research in the development of AI,* in *Synergy - DRS International Conference*, 2020, 1320-1321.

[76] European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, Brussels, 2021, Recital 71 and 72; 'First regulatory sandbox on Artificial Intelligence presented' https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented accessed 17 April 2023.

3.4. *Data Protection Impact Assessment and Algorithmic Impact Assessment*

Data Protection Impact Assessments (hereafter DPIAs) constitute a requirement to be fulfilled by data controllers when the processing of personal data using new technologies is likely to result in high risks to the rights and freedoms of individuals. DPIAs should include the envisaged measures and safeguards to address the risks to rights and freedoms of data subjects (Article 35 of the GDPR). In this regard, the implementation of an AI risk-assessment tool in the public sector would be considered as a new technology that would process personal data and pose high risks to citizens' fundamental rights. Hence, DPIAs should be conducted before implementation to identify the risks of the tool, to reflect on how to tackle them, and to select appropriate mitigating measures to those risks. Therefore, a similar exercise to the one conducted in this paper.

In the case of SyRi, the Court concluded that the only existing DPIA was delivered before the GDPR entered into force and that such assessment was not done for each of the five projects carried out under SyRi legislation.[77] Moreover, it is clear that the principle of transparency towards individuals was not properly addressed in that DPIA since the Court considered that this principle was "insufficiently observed in the SyRi legislation" and that "in no way provides information on the factual data that can demonstrate the presence of a certain circumstance, in other words which objective factual data can justifiably lead to the conclusion that there is an increased risk".[78]

Similarly, in the field of AI, Algorithmic Impact Assessments (hereafter AIAs) are deemed necessary to evaluate the potential impact of algorithmic systems before their deployment. In this regard the European Parliament Research Services opines that ADMSs should not be implemented without a prior AIA except when it is certain that they will not have a significant impact on individuals' lives.[79] The AINow Institute elaborated a report on AIAs where it stressed that the benefits of conducting such an assessment can help identify the "potential issues of inaccuracy, bias and harms to affected communities" and determine possible ways to address these impacts while involving affected community members in that process.[80] In this regard, the Government of Canada has released a Directive on Automated Decision Making which requires the completion of an AIA prior to the production of any ADMSs. In concrete, an AIA risk-assessment tool has been developed to score the impact level of ADMSs. [81]

At the EU level, the proposal for an AI act includes the obligation to conduct a "conformity assessment" to demonstrate compliance with the requirements for high-risk AI systems (e.g., transparency, record-keeping), but there is no reference to an instrument like the AIAs.[82] Although conducting conformity assessments is a frequent scheme for the placement of products in the market in the EU, the author considers that in the case of AI systems this would not suffice, especially because this is a one-time assessment. Apart from carrying out a DPIA, it would be necessary to conduct an AIA which is made publicly available to increase transparency and explainability, and to allow citizens to exercise their rights.[83] This should be included in the public database proposed in the AI act in order to show the risks of the AI system and the measures taken to address them. Indeed, this is also remarked by the EDPB and the EDPS in their Joint Opinion: "this database should be taken as an opportunity to provide information for the public at large on the scope of application of AI system and on known flaws and incidents that might compromise their functioning and

[77] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 paragraphs 6.103-6.105 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[78] The Hague District Court, C/09/550982/HA ZA 18-388 Judgment of 5 February 2020 paragraph 6.87 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 accessed 17 April 2023.

[79] European Parliament Research Services (Panel for the Future of Science and Technology), *Understanding algorithmic decision-making: Opportunities and challenges*, 2019, 88.

[80] D. Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, in *AI NOW Institute*, 2018, 9.

[81] Canada's Directive on Automated Decision-Making: www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 accessed 30 April 2021. Canada's AIA risk assessment tool: www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html accessed 17 April 2023.

[82] Articles 3(20) and 43 Proposal for a Regulation of AI.

[83] In this same regard, *see* Algorithmwatch, 'Civil society open letter demands to ensure fundamental rights protections in the Council position on the AI Act https://algorithmwatch.org/en/fundamental-rights-protections-in-the-council-position-on-the-ai-act/ accessed 17 April 2023.

the remedies adopted by providers to address and fix them".[84]

It is worth noting that, *in casu*, both DPIAs and AIAs would be helpful instruments to enforce compliance with transparency, explainability or unbiased requirements. However, the AINow Institute draws a distinction between them. First, while DPIAs are not shared with the public and apply to public and private organisations, AIAs are designed to engage with affected individuals, researchers, and policymakers. [85] In this regard, it seems that AIAs offer a broader scope for action than DPIAs since they may allow the participation of the ultimate users of the technology in the participatory design of the tool (see sub-section 3.3). In any case, this does not mean that DPIAs should not be conducted (indeed, they still constitute a legal obligation under Data Protection legislation), but in the field of AI, AIAs would complement DPIAs.

3.5. *Audits*

Audits are conducted in the field of data protection to assess a specific organisation's compliance with data protection legislation and to verify that appropriate safeguards are in place when personal data are being processed. Audits are included among the tasks of data-protection officers and supervisory authorities to monitor compliance with the provisions of the GDPR (Articles 39 and 57 of the GDPR). As the ICO indicates, audits are intended to be educative and not punitive. Their objective is to identify weaknesses, risks, or deficiencies in the processing practices in order to encourage compliance with data protection legislation.[86] In this vein, they would constitute a useful instrument to assess whether data protection and privacy by design choices are still valid once risk-assessment tools are implemented in the public sector. Moreover, an audit could focus on assessing AI systems for bias. In fact, in some Member States, like the Netherlands, AI systems used by government agencies have already been audited and assessed on their performance.[87]

In this regard, the EDPB and the EDPS, in their Joint Opinion on the AI Act, consider that high-risk AI systems shall be audited by a third party before obtaining the CE marking that would allow providers to place the product in the market.[88] In fact, it would be highly recommended that the results of periodic audits be registered in the public database included in the proposal for an AI act. According to Articles 51 and 60 of the Proposal, this register would contain information on the algorithmic tool e.g., description of the intended purpose of the AI system, contact details of the provider or a copy of the declaration of conformity assessment.[89] This would facilitate auditing tasks, but it would also enhance transparency. In this regard, as already indicated, for the sake of transparency, it would be advisable that an explanation of the model (logic involved behind it) as well as the results of any AIA and DPIA, be also included in the register.[90] The latter is lacking in the AI act.

In light of the above, a register that contains meaningful information on the AI system appears to be a right approach towards the aim of transparency and even explainability. If the register already includes information on the deployed model and the logic behind its reasoning, citizens could get acquainted with the system before being subject to it. Moreover, this register would allow stakeholders (e.g., lawyers, bar associations, data protection professionals) to provide feedback on the tools' design and hence contribute to build human-centric AI risk-assessment tools.[91]

4. *Conclusions*

The use of AI is on the rise in different

[84] EDPB-EDPS, *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, 2021, 20.
[85] D. Reisman et al., *Algorithmic Impact Assessments*, 7.
[86] Information Commissioner's Office, *A guide to ICO audits*, 2018, 11.
[87] An audit of 9 algorithms used by the Dutch Government, https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government accessed 17 April 2023.
[88] EDPB-EDPS, *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, 10-11.
[89] Article 51 Proposal for a Regulation of AI and Annex VIII Proposal for a Regulation of AI.
[90] AlgorithmWatch, *Automating Society Report*, 2020 https://automatingsociety.algorithmwatch.org, 11; F. Reinhold and A. Müller, *AlgorithmWatch's response to the European Commission's proposal regulation on Artificial Intelligence – A major step with major gaps*, 2021 https://algorithmwatch.org/en/response-to-eu-ai-regulation-proposal-2021 accessed 17 April 2023.
[91] The city of Amsterdam is currently developing an Algorithm Register where citizens will be able to provide feedback and hence participate in building human-centric algorithms in Amsterdam (City of Amsterdam Algorithm Register Beta https://algoritmeregister.amsterdam.nl/en/ai-register accessed 17 April 2023).

sectors and industries of society, including the public sector. This is evidenced by the recent legislative proposals issued by the European Commission in the context of AI, but also by the examples of AI systems and AI projects already tested in the public sector (e.g., SyRi or the Polish unemployment tool). These proposals show the efforts undertaken at EU level to ensure an adequate balance between stakeholders´ economic, commercial, or societal interests, and individuals´ fundamental rights. The implementation of these technologies can be beneficial for our society, but they threaten citizens' fundamental rights like the right to non-discrimination or the right to privacy and data protection. The latter has been the focus of this paper due to the vast amount of personal data usually processed by these AI systems. Therefore, this has triggered an analysis of the risks presented by AI systems and how they can be mitigated by looking at the current EU Data Protection framework. From this, it has been possible to extract some lessons on how to implement this technology by public administrations.

This contribution has identified three main risks regarding the use of AI systems: risk of discrimination, risk of AI systems' opacity, and risk of falling into automation. As it has been observed, the principles of transparency and explainability allow to address the risk of opacity and the risk of biased outcomes. In this regard, the data protection framework currently offers a variety of instruments and principles which strengthen the position of data subjects when confronting those risks: transparency, data protection by design, DPIAs and audits. However, throughout this paper, it has been noticed that the existing legislation requires to bear some considerations in mind in light of AI.

Firstly, to address the 'black-box' issue, a constructive approach where an inherently interpretable model is built would be recommended. This way the AI model would be transparent, and it would be possible to provide an *ex-ante* and *ex post* explanation of the process carried out by the tool. This explanation should include information on the specific decision reached by the tool (e.g., reasons behind the decision) and on the system functionality (e.g., quality of the data, design choices). Secondly, data protection by design mandates that transparency be embedded into the tool during its design process. To this aim, a human-centric AI approach where stakeholders and citizens participate in the design of the technology would be suitable to enhance transparency. Equally, citizens should be informed about the results of DPIAs and AIAs. All design choices and the results of the impact assessments should be documented and included in a publicly accessible register where citizens could provide feedback. These considerations are not fully covered by the proposed AI act.

In light of the above, the study conducted in this article has served to define some general requirements and considerations which can be taken into consideration when implementing a decision-making AI system in the public sector, but which can also be applicable to other sectors and domains. Indeed, the use of AI can directly or indirectly cause a significant impact on individuals' fundamental rights (this is the case of Dutch citizens' who were subject to SyRi's risk assessment). Therefore, it is required that those AI systems be transparent and explainable. In this regard, the AI act has proposed important measures in this field, but adequate requirements of transparency and explainability towards individuals are still lacking. In the author's view, there is still work to do to achieve the aim of a true human-centric AI.

With regard to the above, the AI act envisages the creation of an EU database where high-risk AI systems will be registered. However, the information to be contained in that register would be insufficient to achieve explainability. In addition to the certificate of conformity assessment, the register should contain the results of AIAs and DPIAs and meaningful information on the logic behind the tool. This will show the public what risks were identified and what measures were taken to prevent them. Moreover, this could engage citizens and relevant stakeholders in the process of designing or improving the AI system (participatory design). Indeed, a conformity assessment is not sufficient because although it would show that users and producers of AI products are compliant, it would not provide individuals with meaningful information on the functioning of the tool. These design options and choices can be tested through the establishment of regulatory sandboxes by public authorities which offer a safe environment to experiment with this technology before putting it into service.

To conclude, this research demonstrates that for the implementation of human-centric AI systems respectful with fundamental rights in

the public sector, it is necessary to ensure transparency during the whole process of designing, building and deployment of the technology. Transparency enables explainability, and explainability allows citizens to exercise their rights. This is why the organisation of public campaigns, conferences or participatory sessions with relevant stakeholders and citizens at the initial phases of an AI project can be helpful to build citizen-centric AI. Moreover, a human-centric AI approach requires continuous human oversight and especially in the public sector, it requires that humans are not replaced by machines, nor that they fall into automation. It remains to be seen how the final AI act will look like, but some remarks have already been included in this contribution. Some guidance can be found in the existing EU Data Protection framework.