Transparency and Bias Control in the Judicial Application of Artificial Intelligent. The Case of Spain.

María Pérez-Ugena
(Associate Professor of Constitutional Law – Rey Juan Carlos University[1])

Abstract

Lo studio è volto ad indagare il crescente interesse in merito all'impatto che le decisioni cd. automatizzate potrebbero avere sui diritti individuali e sull'equità. Nello specifico, ci si soffermerà sulle pronunce rese, mediante sistemi di intelligenza artificiale, da organi giudiziari, evidenziando la rilevanza, in tal senso, del diritto alla trasparenza e di quello alla «spiegabilità», che garantiscono la possibilità per gli utenti di comprendere il processo decisionale ed, eventualmente, di contestarne gli esiti. L'analisi che ci si propone di svolgere con riguardo agli algoritmi utilizzati nel settore d'interesse tenderà all'identificazione dei potenziali pregiudizi e ad eliderne l'influenza sulla decisione finale, sottolineando le loro intrinseche criticità soprattutto rispetto all'ingiustizia della medesima. Inoltre, si proverà ad esaminare con attenzione le iniziative dell'Amministrazione spagnola sul punto, focalizzandosi sui limiti etici e legali necessari per salvaguardare i diritti individuali e preservare una giustizia equa.

This study explores the emerging concern about the impact of automated decisions on individual rights and judicial equity. Specifically, it focuses on decisions made by judicial entities and artificial intelligence systems, highlighting the importance of rights to transparency and explainability. These rights ensure that people understand the decision-making process and could challenge or restrict such decisions. A critical analysis of algorithms is carried out to identify and mitigate potential biases, emphasizing the inherent dangers that could lead to unfair resolutions. In addition, current initiatives in the administration of justice in Spain are examined, emphasizing the ethical and legal limits necessary to safeguard individual rights and preserve equitable justice.

[1] Work carried out within the research project: The freedom of expression of the judge in the framework of respect for the principles of judicial ethics in Spain pid2021-127122nb-i00.

1. Introduction

The purpose of this work is the study of transparency and explainability, on the one hand, and bias and its forms of control, on the other, as determining elements for artificial intelligence to have a certain impact in the legal world, especially in the actions of the Judicial Power.

For these purposes, the concepts of transparency, explainability, bias, and control of artificial intelligence (AI) will be defined, along with their implications. It is argued why these elements are important to ensure trust, responsibility and control of AI in the judicial sphere, and what risks they pose. The understanding of bias and control of AI is also presented, emphasizing their importance in ensuring fairness, impartiality, and legal security of AI in the judicial domain.

The main cases and examples of AI application in the legal world, both nationally and internationally, will be analyzed, assessing their degree of transparency, explainability, bias, and control. Additionally, the functioning of these systems is described, and measures that have been or should be adopted to ensure their transparency, explainability, bias mitigation, and control are discussed.

Finally, the current Spanish regulation is examined, and some recommendations and best practices are proposed to achieve transparent, explainable, unbiased, and controlled AI in the legal realm, from both technical and ethical-legal perspectives.

To address this question, a literature review of the main academic, legal and technical sources on the topic is carried out. The benefits and challenges of applying AI in the judicial field are analyzed, as well as the ethical and legal principles that should govern its use. It is concluded that AI can have a complementary and not substitutive role in judicial decisions, as long as certain conditions are met and mechanisms of control and accountability are established.

2. Transparency and explainability of artificial intelligence

AI systems must be transparent and explainable. This means that it should be possible to understand and communicate how they work, what data and algorithms they use, how they make decisions, and what consequences they have. These aspects are fundamental to ensure respect for fundamental rights in the use of AI, and they are rights in themselves. Transparency and explainability should not only be a concern after implementing AI systems but should be integrated from their design and development[2]. This involves adopting a holistic and multidisciplinary perspective that engages all AI stakeholders, from developers and providers to regulators and users. It also involves applying criteria and standards of quality, verification, and validation to ensure the reliability and auditability of AI systems[3].

The General Data Protection Regulation (GDPR) contains a significant set of rules on algorithmic accountability, imposing transparency, processes, and supervision in the use of AI. The GDPR is an undeniable cornerstone in the direction of the new privacy era, setting a minimum standard for rights related to transparency and explainability. These rights apply to decision-making through complex algorithms or AI[4].

The lack of transparency in AI results may be due to the complexity, opacity, or uncertainty of the algorithms, data, or models that the machine uses to perform a task. Transparency is an objective concept that can vary in content depending on the intended purpose. In the context of AI, it is considered both an obligation for various subjects and a subjective right with multiple contents. Furthermore, it is configured as attributions of subjects overseeing AI systems. Given its instrumental nature, transparency is not a static concept, and its content evolves according to the purpose and context in which it is applied.

[2] W. Arellano Toledo, *El derecho a la transparencia algorítmica en big data e inteligencia artificial*, in *RGDA Iustel*, no. 50, February 2019, available at: https://www.iustel. com/v2/revistas/detalle_revista.asp?id=1.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, *A Survey of Methods for Explaining Black Box Models*, in *ACM Computing Surveys*, 2018, vol. 51 (5), 93; L. Cotino Hueso, J. Castellanos Claramunt, *Transparencia y explicabilidad de la Inteligencia Artificial*, Valencia, 2022.

[4] G. De Minico, *Towards an "Algorithm Constitutional by Design"*, in *Biolaw journal*, 2021, 1, 381- 403.

Transparency and explainability are essential to guarantee the right to effective judicial protection, enshrined in Article 24 of the Spanish Constitution[5]. This right not only implies that individuals have the right to access justice, obtain a resolution based on the law, be heard, and defend their legitimate interests but also the right of data subjects not to be subjected to a decision "based solely on automated processing". Article 22 of the GDPR could mean both a right to object to such decisions and a general prohibition on decision-making based solely on algorithms. However, this article applies only when the decision is based "solely" on algorithmic decision-making and applies only when the decision produces "legal effects" or "significantly affects" the individual. Additionally, Article 22 establishes the right of the data subject to human participation in algorithmic decision-making, implying that there must be a person who can listen to and address the objections of the data subject and modify the initial automated decision if it was unfair, biased, or incorrect[6]. This right ensures that no person is subjected to important or significant decisions that affect their rights and daily life when these decisions are based solely on automated processes, through algorithms or computer systems, without direct human intervention[7].

It should, therefore, refer to situations where automated decisions can result in the denial of a right or the loss of a specific opportunity. For example, if an algorithm decides that a person is not eligible for a loan, this automatic decision can deprive that person of the right to access that credit. Even when automated decisions do not have direct legal consequences, they can have a significant impact on a person's life.

The GDPR includes a series of scenarios in which this prohibition does not apply, which are as follows:

First, necessary for the conclusion or performance of a contract between the data subject and a data controller: It must be demonstrated that the processing is necessary for the execution of a contract, and this necessity must be interpreted strictly. This could occur in cases where developers hire individuals to use their personal data during the system training process. Similarly, the data processing

[5] Recommendation on the Ethics of Artificial Intelligence, adopted on November 23, 2021, by UNESCO.

[6] *Ibidem.*

[7] A. Palma Ortigosa, *Decisiones automatizadas y protección de datos personales. Especial atención a los sistemas de inteligencia artificial*, Madrid, 2022.

controller, providing a service to third parties, could use the data of these parties in the context of the service contract they offer[8].
Second, authorized by Union or Member State law applicable to the data controller and providing appropriate measures to safeguard the rights and freedoms and legitimate interests of the data subject: It is crucial to note that the last two legal bases must be established through EU or Member State law, determining the legal basis for processing. In other words, a data controller cannot claim reasons, such as public interest, unless established in an appropriate-level norm. These exceptions may include situations where automated processing is necessary to prevent fraud, tax evasion, maintain the security and reliability of services.

Third, based on the explicit consent of the data subject[9]: the consent of data subjects, which is any freely given, specific, informed, and unambiguous indication of the data subject's wishes, by which they accept, whether by a statement or a clear affirmative action, indicating the acceptance of the processing of personal data concerning oneself. Explicit consent is an exception to the prohibition of automated decisions and profiling, as established in Article 22, paragraph 1. It is important to remember that consent is not always an adequate basis for data processing. In all cases, individuals must receive sufficient information about the intended use and potential consequences of processing so that the consent they provide is an informed choice.

In any case, appropriate measures must be taken to safeguard the rights, freedoms, and legitimate interests of the data subject, including at least the right to obtain human intervention from the con-

[8] See: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: to foster a general understanding of AI systems; to make stakeholders aware of their interactions with AI systems, including in the workplace; to enable those affected by an AI system to understand the outcome; and to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

[9] EU Charter of Fundamental Rights enshrines the protection of personal data as a fundamental right under Article 8, distinct from the respect for private and family life as outlined in Article 7. Article 8 establishes the requirement for a lawful basis for processing. Specifically, it stipulates that personal data must be processed "on the basis of the consent of the data subject or some other legitimate basis laid down by law".

troller, to express their point of view, and to contest the decision. Additional protection is provided for special categories of data.

A right derived from the necessary transparency and explainability of AI systems is also recognized, the right to be fully informed when a decision is based on AI algorithms or made with their participation, especially when it affects their security or human rights. Articles 13 and 14 of the GDPR recognize this right in relation to "the existence of automated decision-making". It must include all information that allows traceability, meaning that the data subject can reconstruct the decision that affects them. To achieve this, it is necessary to inform them about the characteristics and specific logic of the algorithms used, without trade secrets being a pretext to deny any information to the data subject or the judge[10].

From the lack of transparency, significant issues arise regarding trust, verification, and questioning of decisions made by AI systems. Trust implies that people have a reasonable expectation that AI systems[11] will act consistently, predictably, and beneficially for them and for society. Verification implies that people can check and validate that AI systems meet established requirements and standards. Questioning implies that people can object and oppose decisions made by AI systems, especially when they consider them unfair, incorrect, or harmful[12].

Obvious problems can arise, for example, with judicial prediction systems that use AI to analyze data from previous legal cases and estimate the probability that a judge or court will issue a particular judgment or resolution. These systems can be useful in guiding lawyers, judges, and parties in the judicial process but can also pose risks of bias, lack of transparency, and violation of professional secrecy and judicial independence[13].

Similarly, facial recognition systems, understood as AI systems that identify or verify a person's identity from an image or video of their face, can be useful for crime prevention and investigation but can also

[10] M. E. Kaminski, *The right to explanation, explained*, in *Berkeley technology law journal,* 2019, vol. 34 (1), 189-218.

[11] K. Mcgrath, *Accuracy and Explainability in Artificial Intelligence: Unpacking the Terms*, Forty-Second International Conference on Information Systems, Austin, 2021, available at https://bura.brunel.ac.uk/bitstream/2438/26392/4/FullText.pdf.

[12] See Recommendation on the Ethics of Artificial Intelligence, adopted on November 23, 2021, by UNESCO.

[13] C. Poirson, *The Legal Regulation of Facial Recognition*, in K. Miller, K. Wendt (edited by), *The Fourth Industrial Revolution and Its Impact on Ethics*, Cham, 2021, 283-302.

pose risks of invasion of privacy, discrimination, and abuse of power[14]. Moreover, legal assistance algorithms, which provide information, guidance, or legal advice to individuals needing to address a legal issue, can be useful in facilitating access to justice but may also pose risks of lack of quality, responsibility, and protection of personal data. Transparency and explainability in artificial intelligence (AI) systems should not be understood as absolute and uniform concepts but rather as concepts that must adapt to the context and purpose of each AI system. Not all AI systems require the same level of transparency and explainability, and not all users need the same amount and quality of information. Therefore, a balance must be struck between transparency and explainability and other legitimate values and interests. In particular, trade secrets may limit access to the source code of algorithms. The goal is to achieve balanced measures that respect both the privacy of individuals and the innovation of companies.

Although the Regulation has established a framework of rights around the use of automated decisions, it is not as prescriptive as it should be and leaves many issues open to interpretation. In fact, its text resembles more a Directive than a Regulation[15].

Finally, notable differences exist between European and American systems regarding this issue. It is noteworthy how the EU's choice is grounded in democracy and connects with citizens' right to know and control public power by giving greater weight to the right to algorithmic explanation. In contrast, in U.S. law, the lack of regulation has led to the protection of trade secrets prevailing over the right to knowledge, favoring private companies[16].

[14] T. Madiega, H. Mildebrath, *Regulating facial recognition in the EU*, 2021, available at https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/698021/EPRS_IDA%282021%29698021_EN.pdf

[15] See https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: to foster a general understanding of AI systems; To make stakeholders aware of their interactions with AI systems, including in the workplace; To enable those affected by an AI system to understand the outcome, and; To enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

[16] G. De Minico, *Towards an "Algorithm Constitutional by Design"*, cit., 381-403.

3. Bias and Discrimination in AI

The use of algorithms in various fields of society poses important ethical and legal challenges, among which the problem of algorithmic biases stands out. Algorithmic biases are systematic distortions that affect the representation, processing or output of data, and that can generate unfair, inaccurate or discriminatory results. Algorithmic biases can enhance existing or historical biases, which stem from discriminatory attitudes, but are not evident in the applications. They can also create new biases, which arise from the limitations or decisions of the developers, providers or users of the algorithms[17].

The control of biases in AI applied to the judicial system refers to the prevention, detection, and correction of potential unfair, discriminatory, or harmful effects that AI systems may have on the rights and guarantees of individuals involved or affected by judicial processes, such as defendants, victims, witnesses, jurors, etc. Bias control is an ethical and legal requirement aimed at ensuring the fairness, impartiality, and legal certainty of AI in the judicial context. The introduction of potential discrimination in a system can be due to various factors, such as ethnicity, economic status, gender, age, demography, religion or others. These factors can negatively affect minorities or groups underrepresented in the data that are used as reference in computational learning. Data are the key element for the functioning of algorithms, as they determine the quality, accuracy and relevance of the results. However, data are not neutral or objective, but rather reflect the characteristics, values and interests of those who select, collect, label, analyze or interpret them.

This issue is especially worrisome because algorithms, unlike humans, do not have the ability to consciously counteract the biases that they may have incorporated, either consciously or unconsciously, by their developers. Algorithms can act in an opaque, complex or unpredictable way, hindering the understanding, explanation and justification of their results. This can affect the transparency, accountability and trust of algorithmic systems, as well as the rights

[17] See L.H. Nazer , R. Zatarah, S. Waldrip, J.X.C. Ke, M. Moukheiber, A.K. Khanna, R.S. Hicklen, L. Moukheiber, D. Moukheiber, H. Ma, P. Mathur, *Bias in artificial intelligence algorithms and recommendations for mitigation*, 2023, available at https://doi.org/10.1371/journal.pdig.0000278.

and freedoms of the people who use them or are affected by them[18]. Bias implies a clear risk of discrimination, referring to purely formal equality versus discrimination on grounds of birth, race, sex, religion, opinion or any other personal or social condition or circumstance. Situations of direct discrimination can occur, when the rule or decision treats differently and unfavorably a person or certain groups or collectives for any of these reasons. For example, direct discrimination on grounds of sex occurs when it is based on sex or on some characteristic related to it[19].

Situations of indirect discrimination can also occur, when the action or rule, without having a discriminatory appearance, produces disproportionately unequal effects for one of the groups[20]. For example, indirect discrimination on grounds of sex occurs when a criterion that seems neutral is applied, but that affects women more negatively than men[21].

A bias represents an inappropriate deviation in the inference process. Biases are especially problematic when they lead to discrimination in favor of one group to the detriment of another. This phenomenon is not exclusive to artificial intelligence systems, but is inherent to any decision-making process, whether executed by human beings or automatically[22].

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, *A Survey on Bias and Fairness in Machine Learning*, 2019, available at https://doi.org/10.48550/arXiv.1908.09635.

[19] According to Directive 2002/73/EC of the European Parliament and of the Council, direct discrimination is defined as "situation in which a person is, has been, or could be treated less favorably than in a comparable situation on the grounds of sex".

[20] The case mentioned in Judgment Spanish Constitutional Court n. 145/1991 is illustrative, referring to the situation of female cleaners in a public hospital who were receiving a lower salary than male laborers. The doctrine established by the Constitutional Court from this Judgment prohibits the unequal valuation of equivalent jobs when this differential treatment is based on the gender of the workers. This implies that, considering that the majority of women occupy the position of cleaners and the majority of men occupy the position of laborers in this specific case, providing a lower salary to cleaners is, in fact, prejudicing women. Even when the harm occurs indirectly. In such cases, the Constitutional Court has understood that the difference becomes suspect unless it is justified that it is not based on gender but on the characteristics of the work. Hence, equal pay is not only required for the same job but also for a different job of equal value (in Judgments nn. 198/1996; 240/1999).

[21] R. Serra Cristobal, *La discriminación indirecta por razón de sexo*, in M.J. Ridaura Martinez, M.J. Aznar Gomez (coord. by), *Discriminación versus Diferenciación (Especial referencia a la problemática de la mujer)*, Valencia, 2004, 365-398.

[22] E. Ferrara, *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and*

Algorithmic biases can be of different types, depending on the origin, level or effect of the distortion. For example, one can distinguish between input, process or output biases, depending on whether they affect the data, the algorithms or the results.

These types of biases are related to the stages of the life cycle of an artificial intelligence system, from the collection and selection of data to the design and evaluation of algorithms, through the training and validation of models, and finally the implementation and use of the results.

Each of these stages can introduce errors or deviations that affect the quality, accuracy, fairness and transparency of the system. In addition, one can distinguish between technical, cognitive or social biases, depending on whether they come from the limitations or decisions of the developers, the users or the contexts of application. Technical biases refer to the problems derived from the complexity, opacity or inadequacy of the algorithms or data. Cognitive biases refer to the prejudices or heuristics that influence the reasoning or behavior of the users. Social biases refer to the inequalities or discriminations that affect certain groups or collectives for reasons of gender, race, age, religion, etc.

The introduction of biases can occur in two fundamental ways. First, by selecting data that does not faithfully represent reality, a phenomenon known as sampling bias. A concrete example would be the preference for using more images of white faces than of other colors, thus generating a distortion in the representation of ethnic diversity.

Second, bias can emerge by reflecting prejudices already existing in the training data of the algorithm. For example, the use of historical information on hiring decisions that favored men over women could lead the algorithm to learn and perpetuate that discrimination, as happened in the case of Amazon[23].

mitigation strategies, 2023, available at https://arxiv.org/pdf/2304.07683.pdf.

[23] This case is an example of how artificial intelligence (AI) can replicate and amplify human biases if measures are not taken to prevent it. Amazon's recruitment system, which relied on a machine learning algorithm, showed a preference for male candidates and penalized women aspiring to technical positions. This was because the algorithm learned from the company's historical hiring data, which was predominantly male-oriented. The algorithm also assigned a lower score to resumes containing the word "woman" or the names of women's colleges. Amazon attempted to correct the algorithmic bias, but ultimately decided to abandon the project due to a lack of confidence in its neutrality. See: https://www.bbc.com/mundo/noticias-45823470.

It is important to emphasize that these biases are not mere technical errors; they have relevant implications. It has been proven that they often reproduce and perpetuate, reinforcing dynamics of domination, privilege, and discrimination. This increases the risk that existing inequalities may be exacerbated and solidified through automation and machine learning[24].

Bias manifests in the imbalance in datasets due to the presence of underrepresented groups. A clear example of bias is found in studied datasets that exhibit a noticeable imbalance, as over eighty percent of the reference subjects in the databases have light skin. This implies an overwhelming majority of individuals with this skin tone compared to other shades, indicating an evident imbalance towards this particular group[25].

The implication of this imbalance is that when analyzing or using these datasets, the information and conclusions obtained may be biased towards the majority representation of individuals with light skin. This underscores the importance of considering and addressing these imbalances to ensure that the analysis results are not skewed by the lack of adequate representation of diverse groups. This disparity can bias the analysis, favoring light-skinned groups and marginalizing those with dark skin. Additionally, it is necessary to consider not only genders but also subdivisions by race to achieve a more accurate representation[26].

The influence of bias also affects widely used datasets in machine learning, such as ImageNet and Open Images. The representation biases in these datasets are evident, and there has been advocacy for the inclusion of geographical diversity as a mitigation measure. In the field of Natural Language Processing (NLP), representational biases are identified in knowledge bases used in various applications[27].

Several cases have been raised where courts have ruled on whether there is bias in a particular situation.

The Syri Program (Risk Indication System) is a system used in the Netherlands within the context of the Implementation and Income Structure Organization Act (SUWI), specifically under Article 65.2.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, *A Survey on Bias and Fairness in Machine Learning*, cit.

[25] *Ibidem.*

[26] *Ibidem.*

[27] *Ibidem.*

Its main purpose was to assess the risk that an individual or legal entity could illegally use government funds related to social security. The goal was to identify potential cases of fraud or misuse of these funds. To carry out this risk assessment, the system operated in two main phases[28]:

In the first phase, all relevant data about the individual or entity in question are collected. This data may include information about scholarships, insurance, employment, sanctions, place of residence, and other aspects related to the social security benefits received. This data undergoes a pseudonymization process, meaning that direct personal identifiers are removed, and a unique number is assigned to each dataset. Once the data has been pseudonymized, predefined models and risk factors are used to analyze the collected information. These models and risk factors assign a risk rating to each dataset based on the probability of illegal use of government funds. If the resulting risk rating is high, the process proceeds to the second phase.

In the second phase, cases that have received a high risk rating undergo additional analysis by a specialized unit. This unit is responsible for a more detailed and definitive assessment and rating of the risk. This involves a more thorough review of the data and a more precise decision-making process regarding the likelihood of fraud or misuse of government funds.

In summary, the Syri Program is a system used to assess the risk that individuals or entities are illegally using government funds related to social security in the Netherlands. The process involves data collection, pseudonymization, the use of models and risk factors, and a two-phase evaluation to determine potential and, ultimately, definitive risk. This could help authorities identify possible cases of fraud and take appropriate actions accordingly[29].

The Judgment of the District Court of The Hague in the Netherlands, dated February 5, 2020, concluded that the SyRI algorithmic

[28] N. Appelman, R. Ó Fathaigh, J. van Hoboken, *Social Welfare, Risk Profiling and Fundamental Rights: The Case of SyRI in the Netherlands*, in *Jipiteg*, 2021, 12 (4), available at https://www.jipitec.eu/issues/jipitec-12-4-2021/5407; M.J. Rivas Velasco, *Uso* ético *de la inteligencia artificial en justicia*, in *Diario La Ley*, Nº 10327, Sección Tribuna July 13, 2023, 1-19.

[29] N. Belloso Martìn, *La problemática de los sesgos algorítmicos (con especial referencia a los de género). ¿Hacia un derecho a la protección contra los sesgos?*, in F.H. Llano Alonso (dir. by), J. Garrido Martín, R.D. Valdivia Giménez, F.H. Llano Alonso (coord. by), *Inteligencia artificial y filosofía del Derecho*, Murcia, 2022, 45-78.

system used for fraud risk assessment in the field of Social Security and the Ministry of Finance was incompatible with the law. It did not meet the required standards of proportionality and transparency and violated the provisions regarding respect for private life as recognized by Article 8 of the European Convention on Human Rights. The basis of this decision focused on the fundamental concern that the system had the potential to discriminate against the citizenry, relying on algorithms for data analysis and risk assessment. The Court maintained that data processing becomes significantly relevant when its consequences are substantial enough to significantly influence the behavior and decisions of the individuals involved. This influence can manifest through long-term or enduring effects on the affected person, or even trigger their exclusion or experiences of discrimination. The government violated citizens' right to privacy by not disclosing the nature of the algorithms used in the risk assessment model or providing information about the method used in risk analysis. This led to the affected individuals being unaware that their personal information was used for such purposes and lacking the ability to verify the accuracy of the data used. As a result, when configuring risk profiles, there was the potential for involuntary connections based on biases or prejudices.

A second example, occurring in the private sector, of bias resulting from the application of AI systems led to the judgment of December 31, 2020, by the Ordinary Court of Bologna (Italy) regarding the Deliveroo case. The company in question utilized an application called "Frank" to manage the allocation of gastronomic orders. A lawsuit was filed alleging discrimination in the reliability and availability indices used by the algorithm to manage the work sessions of its employees, particularly the delivery riders. This application relied on an algorithm designed following the instructions and guidelines provided by the company itself. The primary goal of this algorithm was to score workers and determine their priority for booking work sessions.

The algorithm was based on two key indicators:

First. Reliability: This indicator focused on how often a worker failed to fulfill the work sessions they had reserved. In other words, it evaluated the worker's punctuality and commitment to the booked sessions. Each instance of non-compliance was recorded and affected the worker's score.

Second. Availability: The second index assessed how often a worker was available during peak demand hours. In this case, particular emphasis was placed on the hours from 8:00 to 10:00 PM on Fridays, Saturdays, and Sundays, when service demand is typically higher.

However, controversy arose because the "Frank" application did not consider, as per the company's instructions, justified absences of workers. This had a direct impact on the riders' scores, penalizing workers in such situations. Justified absences could result from situations such as strikes, illnesses, or other valid and legitimate circumstances preventing workers from fulfilling their work reservations.

Justified absences were treated similarly to unjustified ones, negatively affecting workers' ability to choose their work schedules. This essentially created a discriminatory system where justified absences were unfairly punished, restricting job opportunities for the riders.

Although the legal resolution did not provide a detailed explanation of the internal workings of the algorithm or its level of transparency, it was concluded that the information input into the system was biased by the company that commissioned its development. This bias infringed on the workers' rights in terms of equality and non-discrimination. It was determined to be "indirect discrimination", indicative of "unconscious and deliberate blindness" on the part of the company.

In Spain, the issue was addressed through Law 12/2021, dated September 28, known as the "Riders Law"[30]. One of the most prominent aspects of this legislation is the obligation imposed on these companies to adapt to accurately reflect the employment relationship they have with delivery riders. For this purpose, they are required to share the algorithms they use and must also provide detailed information about the rules on which these algorithms are based. Additionally, they are encouraged to share this information with unions, implying greater transparency regarding decisions that may affect the working conditions of employees.

The Riders Law introduces a new provision in Article 64.4, stating that: "*The works council, with the periodicity that corresponds in each case, shall be entitled to: 'Be informed by the company about the parameters,*

[30] This law introduces modifications to the Consolidated Text of the Workers' Statute Law, which was approved by Royal Legislative Decree n. 2/2015 on October 23, with the primary purpose of guaranteeing labor rights for individuals engaged in the delivery of goods and services through digital platforms, such as Glovo, Deliveroo, or Uber Eats. Principio del formularioFinal del formulario.

rules, and instructions on which the algorithms or artificial intelligence systems that affect decision-making and may impact working conditions, job access, and maintenance, including profiling, are based"[31].

Finally, there is another type of bias that may be even more concerning, and that is bias in the interpretation of artificial intelligence results. This human bias manifests when we uncritically accept the results of an artificial intelligence system as true and immutable, adopting an "authority principle" based on the expectations generated by such systems.

In other words, this bias involves blindly trusting the results of AI without questioning or critically analyzing their validity. This can lead to erroneous or unjust decisions, as the biased interpretation of results may be influenced by unfounded biases or expectations. Consequently, it is essential to address both the inherent biases in data and algorithms and human bias in interpreting results to ensure more objective and equitable decision-making[32].

To address this issue, a multidimensional approach is proposed, combining different strategies and measures to prevent, detect, and mitigate algorithmic biases. These strategies and measures include the ethical and responsible design of algorithms, auditing and assessing their impact, education and training for the involved stakeholders, user and civil society participation and empowerment, and the regulation and oversight of algorithmic systems. It is concluded that a regulatory and ethical framework is necessary to ensure respect for democratic principles and values, as well as the protection of human rights and fundamental freedoms in the use of algorithms.

4. Artificial intelligence in law and enforcement decisions based on automated data processing and their application to the Spanish judicial system.

In the realm of law enforcement, AI systems focus on what we could term "anticipatory intervention". Instead of directly addressing the conventional concept of a crime characterized by a typical,

[31] Ministry of Labor and Social Economy, *Algorithmic Information in the Workplace: Practical Guide and Tool on the Corporate Obligation of Information Regarding the Use of Algorithms in the Workplace*, Government of Spain, May 2022. Available at: https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/trabajo14/Documents/2022/100622-Guia_algoritmos.pdf.

[32] *Ibidem.*

unlawful, and culpable conduct regarding an event that has already occurred, it is oriented towards a preceding logical and temporal phase. In this sense, the goal is to anticipate when a criminal act might materialize and, above all, prevent it from happening. This approach resembles more of a policing function than a judicial process since courts come into play after the events have taken place.

This involves the use of predictive models, raising intense doubts about their legality. It is necessary to determine whether these tools are legally valid and respect fundamental guarantees[33].

The distinction between predictive and investigative tasks of the police necessarily involves working with profiles. Information is obtained to place individuals who might be involved in criminal activities in predictable times, places, and conditions. In this way, an attempt is made to foresee where, who, when, and why certain crimes occur, even taking into account environmental factors.

The Integral Monitoring System for Gender-Based Violence Cases (VioGén), implemented in Spain since July 2007, within the framework of the Spanish law (Organic Law 1/2004, of December 28, "on Comprehensive Protection Measures against Gender Violence")[34].

This system represents a significant advancement in the fight against gender violence in the country. This system operates through a centralized database that stores key information about gender violence cases, shared and managed by various institutions and law enforcement agencies.

The objectives of this system are to bring together the different public institutions that have competencies in the field of gender violence, to add all the information that may be related and of interest, to

[33] The AI solutions for law enforcement and judicial authorities in criminal matters, as highlighted by the European Parliament, must fully respect certain principles. Among these are the principles of human dignity, non-discrimination, freedom of movement, presumption of innocence, and the right to defense, including the right to remain silent, freedom of expression and information, freedom of assembly and association, equality before the law, equal defense, and the right to effective judicial protection and a fair trial, in accordance with the Charter and the European Convention on Human Rights.

The report also calls for a mandatory assessment of the impact on fundamental rights before the implementation or deployment of AI systems in the police or judicial context to assess potential risks to fundamental rights..

[34] See: https://www.interior.gob.es/opencms/es/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen/.

predict the risk to which a woman who has been a victim may be exposed, to carry out both monitoring and protection in an adequate way and to carry out a preventive work, issuing warnings, alerts and alarms, through the Subsystem of Automated Notifications, when it is believed that the integrity of the victim is in danger[35].

With all the information collected and included in the VioGén System, it will assign a level of risk – "not appreciated", "low", "medium", "high" or "extreme" –, which can be modified upwards by the agents if they consider it necessary to better protect the victim. The result is communicated to the court and the prosecutor in an automated report that is included in the police report. Each level of risk entails specific police measures of mandatory and immediate application.

In cases where that risk is "not appreciated", the agents focus on informing the woman of the available resources that she can go to and access. When the system determines a level of "low" risk, the woman will be provided with a permanent contact telephone number and telephone or personal contacts will be made discreetly and agreed with the victim.

They will try to find out through her the judicial resolutions of the case, as they may increase the danger for the woman and require greater protection.

The agents must inform the aggressor that his case is subject to police control; if he has weapons, they will initiate the process to withdraw them; and they will punctually control the penitentiary information in the VioGén System to know their possible exits from prison.

With a "medium" level of risk, the measures increase and it is considered whether it is necessary to admit the victim to a shelter. Occasional controls are established at her home, at work and at the children's schools and the Prosecutor's Office is urged to assign a telematic control device to the aggressor.

With a high level of risk, and if the aggressor is not located, the victim should be invited to go to a shelter or to change her address and the controls on the home or workplace will be frequent. Random checks will also be carried out on the aggressor, contacting people

[35] In the validation phase, it was observed that the tools have a good and similar predictive validity than the other tools of risk evaluation in VCP. VPR 4.0 is sensitive to the detection of the risk of recidivism and presents a probability of risk of detecting false negative of 5,1%. It is a tool capable of detecting those subjects with a low risk of recidivism.

in his environment as well. The protection of the victim in cases of "extreme" risk will be permanent and, if necessary, the entrances and exits of the children at school will also be monitored, in addition to establishing an exhaustive control over the aggressor.

Each victim is also provided with a personalized security plan (PSP) with self-protection measures, such as always carrying a mobile phone, making safe use of social networks, adopting security routines in travel or planning an escape routine in case of new aggression.

In this way, interinstitutional coordination works as a fundamental pillar of the system. VioGén promotes collaboration among various entities, including the police, the courts, the social services and the victim support organizations. This coordination ensures an effective and well-orchestrated response to gender violence situations.

The system is continuously updated to adapt to the changing needs in the fight against gender violence. This flexibility guarantees that VioGén remains an effective and modern tool to address this social problem[36].

One of the most notable features of VioGén is its personalized monitoring approach. This allows individual monitoring of victims, facilitating the implementation of specific measures designed to ensure their safety. When an incident or a risk situation is detected, the system is activated through automated notifications.

These actuarial tools of VioGén support the strong evidence of the need for a plan for the protection of victims according to the risk assessment obtained from each case. Neither VPR nor VPER are designed to evaluate psychological aspects or constructs. The transparency and reliability of these tools are designed to make predictions and identify the subjects with the highest risk of recidivism in order to be able to assign the protection resources in the most efficient way possible.

It is necessary to differentiate between different situations or areas of action when analyzing the possible effects of applying artificial intelligence to the Administration of Justice. Distinctions have been made among procedural processing, criminal investigation, and judicial decision-making, as it is evident that artificial intelligence can

[36] Since the creation of the Integral Monitoring System for Gender-Based Violence Cases in July 2007, a total of 701,563 cases have been analyzed, of which 73,072 (10.4%) remain active. See: https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/violencia-contra-la-mujer/estadisticas/2023/ESTADISTICA-ENERO-2023.pdf

play a very different role in each of these three areas[37].

Firstly, concerning procedural processing, artificial intelligence in this case would be a complement or auxiliary means for the judiciary by automating various facets of the judicial process. Not always does the incorporation of new technologies in the judicial process solve all problems and expedite the administration of justice. While technology can improve the speed and efficiency of justice, it is important to maintain a realistic approach and not overestimate its benefits. The judicial process must continue to respect fundamental principles such as immediacy, contradiction, orality, and publicity[38].

The use of algorithms to predict criminal behaviors and assign risk levels, whether for initial or repeat offenses, poses the inherent risk of intrusive surveillance and affecting fundamental guarantees of society members regarding privacy and individual rights. The automated collection and evaluation of data can have significant implications in terms of freedom and fairness.

Additionally, algorithms can be used to automate various aspects of the judicial process, such as reviewing past cases for consistent jurisprudence, generating reports, and even making judicial decisions. However, this automation carries the risk of affecting fundamental guarantees, as the absence of direct human supervision could result in biased or unjust decisions.

In the Spanish system, the current regulation, in accordance with the Organic Law 7/2021, of May 26, on the protection of personal data processed for the prevention, detection, investigation, and prosecution of criminal offenses and the execution of criminal sanctions, prohibits, in Article 14, decisions based solely on automated individual processing[39]:

[37] C. Lorenzo Pérez, *Inteligencia artificial en la administración de Justicia: regulación española y marco Europeo e Internacional.* Proyectos desarrollados por el Ministerio de Justicia de España. Dirección General de Transformación Digital de la Administración de Justicia. Ministerio de Justicia, 2022, available at https://www.cej-mjusticia.es/sede/publicaciones/ver/13637, follows the classification proposed by A. Del Moral García, which can be seen in *Robotización e Inteligencia Artificial en la Justicia, organizado por el Ministerio de Justicia en colaboración con AMETIC, 16 de Marzo de 2022* (see: https://youtu.be/0S8kfKm8GZI).

[38] C. Mira Ros, *El expediente judicial electrónico*, Ministerio de Justicia, Madrid, 2010.

[39] In the same vein, every data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or significantly affects him in a similar way (Art. 22 GDPR). This right ensures that no individual is subjected to important or significant decisions

"Decisions based solely on automated processing, including profiling, that produce negative legal effects for the data subject or significantly affect them, unless expressly authorized by a law or by European Union law. The enabling regulation for the processing must establish appropriate measures to safeguard the rights and freedoms of the data subject, including the right to obtain human intervention in the review process of the decision taken. 2. The decisions referred to in the preceding paragraph shall not be based on the special categories of personal data referred to in Article 13, unless appropriate measures have been taken to safeguard the rights, freedoms, and legitimate interests of the data subject. Profiling that leads to discrimination of natural persons on the basis of special categories of personal data established in Article 13 is prohibited".

In the Draft Law on Digital Efficiency Measures of the Public Service of Justice, the existing regulation is expected to be expanded to incorporate the possibility of proactive and assisted automated actions, for which the use of artificial intelligence techniques will be allowed. The project distinguishes between automated, proactive, and assisted actions[40].

An automated action is a procedural action produced by a properly programmed information system without the need for the intervention of a natural person in each individual case. The computer systems used in the Administration of Justice will enable, according to the project, the automation of simple procedural or decision-making actions that do not require legal interpretation. These include numbering or paging of files, referring matters to the archive when procedural conditions allow it, generating copies and certificates,

that impact their rights and daily life when these decisions are solely based on automated processes, through algorithms or computer systems, without direct human intervention. Thus, the denial of a loan, a job offer, the approval or rejection of an insurance application, or the allocation of public resources should not be based solely on algorithms without a person having had the opportunity to influence the decision. It should, therefore, pertain to situations where automated decisions can result in the denial of a right or the loss of a specific opportunity. For example, if an algorithm decides that a person is not eligible for a loan, this automatic decision may deprive that person of the right to access that credit. Even when automated decisions do not have direct legal consequences, they can have a significant impact on a person's life, emphasizing the need to ensure fairness and transparency in an increasingly automated world.

[40] Draft Law on Digital Efficiency Measures for the Public Judicial Service, which transposes into the Spanish legal system Directive (EU) 2019/1151 of the European Parliament and of the Council of June 20, 2019, amending Directive (EU) 2017/1132 with regard to the use of digital tools and processes in the field of Company Law. See: https://www.mjusticia.gob.es/es/AreaTematica/ActividadLegislativa/Documents/APLEficienciaDigitalAudPubeinformes_actual.pdf.

generating books, checking representations, and declaring finality, in accordance with procedural law[41].

Proactive actions are a type of automated actions that self-initiate by information systems without human intervention. These actions leverage the information incorporated into a file or procedure of a public administration for a specific purpose, to generate notices or direct effects for other purposes, in the same or other files, of the same or another public administration, in any case in accordance with the law. It is expected that the State Technical Committee for Electronic Judicial Administration will collaborate with other public administrations to identify processes that can be proactive and define the necessary parameters and compatibility requirements for them.

Assisted actions, finally, aim to facilitate the work of justice professionals, streamline processes, and improve the quality of resolutions. An assisted action is one for which the information system of the Administration of Justice creates a draft of a complex document, in whole or in part, from data that can be generated by algorithms. This draft can serve as a basis or support for a judicial or procedural resolution but does not have validity on its own without the approval of the competent authority. Justice Administration systems ensure that the user can request, modify, and reject the documentary draft to their liking. For the draft to become a judicial or procedural resolution, it is necessary for the Judge, Magistrate, Prosecutor, or Legal Officer, as appropriate, to validate the final text and identify, authenticate, or electronically sign it as established by law, in addition to complying with the requirements that procedural laws demand.

The Administration of Justice must ensure in all these actions: a) That it can be identified, tracked, and explained if an action is automated or proactive. b) That the same action can be performed in a non-automated way. c) That automated actions that have already been carried out can be deactivated, undone, or invalidated.

In addition, certain common requirements for automated, proac-

[41] Articles 56 to 58 of the Draft Law on Digital Efficiency Measures in the Public Justice Service, transposing into the Spanish legal system Directive (EU) 2019/1151 of the European Parliament and of the Council of June 20, 2019, amending Directive (EU) 2017/1132 regarding the use of digital tools and processes in the field of company law. https://www.mjusticia.gob.es/es/AreaTematica/ActividadLegislativa/Documents/APLEficienciaDigitalAudPubeinformes_actual.pdf

tive, and assisted actions are demanded. On the one hand, the State Technical Committee for Electronic Judicial Administration will, if necessary, be responsible for defining the specifications, programming, maintenance, supervision, quality control, and audit of the information system and its source code when it comes to an automated, assisted,

Furthermore, with the same guaranteeing spirit, the criteria for automated decision-making will be public and objective in the cases mentioned above, and the decisions made at each moment will be recorded. Finally, the systems will have management indicators determined by the National Judicial Statistics Commission and the Technical State Committee of Judicial Electronic Administration, each within its scope of competence.

Lastly, there is a regulation of automated actions in the field of Justice Administration in Spain through Law 18/2011, dated July 5, which regulates the use of information and communication technologies in the Administration of Justice in Spain. Article 42 of this law establishes that automated actions in the field of Justice Administration must be defined in advance by the Technical State Committee of Judicial Electronic Administration (CTEAJE).

The Technical State Committee of Judicial Electronic Administration (CTEAJE) is the entity responsible for establishing the specifications, programming, maintenance, supervision, quality control, and, where appropriate, auditing of the information systems used in automated actions.

These automated systems must include management indicators defined by the National Judicial Statistics Commission and the Technical State Committee of Judicial Electronic Administration, each within its scope of competence.

In comparative law, a reference is the study of the Loomis case, which was brought before the Wisconsin Supreme Court in 2016. It sheds light on the use of artificial intelligence (AI) systems in the field of criminal justice. In particular, it focused on the COMPAS program (Correctional Offender Management Profiling for Alternative Sanctions)[42], which uses a predictive machine learning algorithm adopted by the United States Justice Administration to

[42] The program is based on a machine learning algorithm that analyzes 137 variables about an individual's profile, such as age, gender, race, criminal history, educational level, drug consumption, or family situation, among others. Based on this data, the algorithm assigns a risk of recidivism score ranging from 1 to 10, with 10 being the highest.

determine the risk of an individual's recidivism at various stages of the criminal process, and how this probability of recidivism can influence the decision of whether a person should be released or remain detained.

This case sparked a significant debate about respecting the "due process" and raised fundamental questions about transparency and fairness in the use of algorithms in the justice system[43]. The judge who issued the verdict partially relied on the risk score of COMPAS, which was 8 out of 10. Loomis appealed the verdict, arguing that the use of COMPAS violated his right to an individualized sentence and to know the evidence against him, thereby violating the right to "due process". The plaintiffs argued that decision-making based on the COMPAS program did not provide individuals with a fair and equitable opportunity to be heard before being deprived of their freedom. However, despite concerns, the deliberate existence of bias in the program due to the racial origin of individuals could not be proven. Nevertheless, evidence was presented indicating that COMPAS consistently yielded a higher percentage of recidivism for people of African descent[44].

Despite everything, the Wisconsin Supreme Court rejected his appeal and endorsed the use of COMPAS as a complementary tool for judicial decision-making, provided that defendants were adequately informed of its limitations and constitutional principles were respected[45].

This case highlighted a critical issue related to the opacity and lack of understanding of how these AI systems operate[46]. Although conclusive evidence of intentional racial discrimination was not found,

[43] M.P. Roa Avella, J.E. Sanabria-Moyano, K. Dinas-Hurtado, *Uso del algoritmo compas en el proceso penal y los riesgos a los derechos humanos*, in *Revista Brasileira de Direito Processual Penal,* 2022, vol. 8, 1, 275-310, available at https://doi.org/10.22197/rbdpp.v8i1.615; L. Martinez Garai, *Peligrosidad, algoritmos y* due process: *el caso* State v Loomis, in *Revista de Derecho Penal y Criminología*, 2018, 3.a Época, n.o 20, 485-502, available at: https://roderic.uv.es/bitstream/handle/10550/72730/135792.pdf?sequence=1&isAllowed=y.

[44] N. Belloso Martìn, *Algoritmos predictivos al servicio de la justicia: ¿una nueva forma de minimizar el riesgo y la incertidumbre?*, in *Revista de la Faculdade Minera de Direito*, 2019, 22 (43), 1-31; D. Bonsignore Fouquet, *Sobre inteligencia artificial, decisiones judiciales y vacíos de argumentación*, in *Teoría & Derecho. Revista De Pensamiento Jurídico*, 2021, 29, 248-277.

[45] A. L. Washington, *How to Argue with an Algorithm: Lessons from the COMPASProPublica Debate*, in *The Colorado Technology Law Journal,* 2019, 17(1), 1-37.

[46] See: https://www.technologyreview.es/s/13800/caso-practico-probamos-por-que-un-algoritmo-judicial-justo-es-imposible.

the apparent disparity in results raised doubts about the fairness of the COMPAS program. This debate underscores the importance of transparency in the operation of algorithms and in decision-making based on them.

It is essential to understand that the reliability and fairness of these systems depend to a large extent on the quality of the data used and the impartiality of the algorithms. In the case of COMPAS, it was discovered that there was a higher probability of false positives for African-American offenders compared to Caucasian offenders. This means that COMPAS was more prone to incorrectly predict that African-Americans had a higher risk of recidivism in the commission of crimes, which could have serious consequences for the affected individuals.

The resolution of the Loomis case, while not confirming intentional racial discrimination, emphasizes the importance of transparency in the operation of these systems. A better understanding of how algorithms assess and score individuals, as well as the data on which they are based, is essential to ensure fairness and justice in the criminal justice system. It also highlights the need to address potential biases in AI systems and establish adequate mechanisms to ensure that decisions based on these algorithms are fair and non-discriminatory.

In conclusion, the Loomis case and the use of COMPAS emphasize the importance of critically addressing the use of algorithms in criminal justice and the need to ensure transparency and fairness in decision-making based on AI systems. The bias in the results of COMPAS underscores the importance of a careful evaluation and consideration of the limitations of these systems to avoid making decisions solely based on their results, especially when individual liberties are at stake.

5. Autonomy, Responsibility and AI Personality

The autonomy of AI refers to the degree of independence and self-control that algorithms have to perform a task without human intervention or supervision. AI autonomy can range from the simple execution of pre-programmed instructions to the generation of its own actions based on learning and adaptation to the environment. Additionally, these elements have an unpredictable nature and the ability to inflict physical harm, marking the beginning of a new phase in the relationship between humans and technology. Unlike software and the Internet, they are designed to interact with the offline world. Their ability to act physically in the real world translates into the potential to cause material harm to people or objects. From these considerations arise important questions about the responsibility, ethics, and safety of decisions made by the machine, as well as the respect for the fundamental rights of individuals affected by such decisions[47].

The autonomy of AI has potential benefits and risks for human society. Some of the benefits include that autonomous AI can enhance efficiency, productivity, and innovation across various sectors and activities. Additionally, it can broaden the capabilities and opportunities for individuals by facilitating access to information, knowledge, learning, communication, and collaboration. This, in turn, translates into an enrichment of diversity and creativity in human culture, giving rise to new forms of expression, art, entertainment, and knowledge.

It poses, likewise, risks, such as physical, moral, or economic harm to individuals or the environment, whether due to accidents, errors, malfunctions, sabotage, or misuse. Secondly, autonomous AI can threaten the privacy, security, autonomy, dignity, and rights of individuals, especially in vulnerable or marginalized groups, by collecting, processing, and using their personal data or influencing their decisions and behaviors. Additionally, autonomous AI can lead to unemployment, precarity, inequality, and social exclusion by replacing or displacing human workers or reducing their incomes, benefits, and opportunities.

Lastly, it is noteworthy that autonomous AI poses the risk of eroding trust, responsibility, transparency, and democracy in society by complicating the understanding, control, supervision, and regulation of its processes and outcomes. To maximize the benefits and minimize the risks of AI autonomy, several measures are required, such as developing and implementing ethical, legal, and technical standards for autonomous AI based on universal values and principles of human rights, democracy, and justice. It involves encouraging participation, collaboration, and dialogue among different stakeholders affected by autonomous AI, including developers, users, regulators, researchers, educators, the media, and civil society.

[47] M. Barrio Andrés, *Delitos 2.0. Aspectos penales, procesales y de seguridad de los ciberdelitos*. Madrid, 2018.

It also entails promoting education, training, awareness, and empowerment of individuals to leverage opportunities and address the challenges of autonomous AI by developing their digital, critical, creative, and ethical competencies. Furthermore, it requires ensuring equity, inclusion, and solidarity in the access, use, and impact of autonomous AI, protecting and supporting the most vulnerable or disadvantaged groups, and reducing social and digital gaps.

One crucial aspect is determining whether AI has legal personality[48]. Identifying a robotic personality is essential to determine the potential civil liability arising from the widespread actions of robots[49]. Thus, if subject to liability, legal consequences for the machine's actions or omissions could be attributed, both in civil and criminal contexts[50]. From this perspective, the existence of two types of AI responsibility can also be considered. Direct responsibility implies that the machine is considered a legal subject, capable of being accountable for its own actions, thereby recognizing its legal personality. Indirect responsibility implies that the machine is considered an object of law, and responsibility falls on the user, programmer, manufacturer, or owner of the machine, depending on the case. Currently, there is no consensus on the most suitable legal framework to regulate AI responsibility[51].

Each type of responsibility has certain advantages and disadvantages, as well as criteria for assigning it in each case. It could be considered whether the direct responsibility of AI requires the machine to have certain cognitive abilities or if it is sufficient for it to have a sufficient degree of autonomy and intelligence. It could also be examined whether the indirect responsibility of AI depends on the nature of the task, the level of risk, the predictability of behavior, or the possibility of controlling the machine. Additionally, it could be assessed whether there are legal, technical, or ethical mechanisms to ensure the transparency, traceability, and accountability of AI, as well as to protect the rights and interests of individuals involved or affected by its decisions. These issues are relevant to the develop-

[48] A. Carrasco Perera, *A propósito de un trabajo de Gunter Teubner sobre la personificación civil de los agentes de Inteligencia Artificial avanzada. (Digitale Rechtssubjekte? Zum privatrechtlichen Status autonomer Softwareagentem", Archiv für die Civilistische Praxis, 218, 2018, pp. 155-205)*, in *CESCO*, January 2019, (See: http://centrodeestudiosdeconsumo.com), 2.

[49] J. García-Prieto Cuesta, *¿Qué es un robot?*, in M. Barrio Andrés (dir. by), *Derecho de los Robots*, Madrid, 2018, 25 ss.

[50] *Ibidem*.

[51] S. Diaz Alabart, , *Robots y responsabilidad civil*, Madrid, 2018.

ment and regulation of AI, as they involve legal, social, and philosophical aspects that must be addressed with rigor and prudence[52].

6. Conclusions

Artificial intelligence poses risks and opportunities in the legal field, especially in judicial decision-making, making it necessary to establish limits on its use.

The cases analyzed regarding the use of artificial intelligence systems in the judicial domain highlight the problems arising from their application, both due to discrimination and the potential lack of transparency and explainability of biased algorithms. This experience serves as a reference to understand the limits being set in legislative provisions in Spain.

Furthermore, the transparency of the algorithms used and the "explainability" of judicial decisions are fundamental aspects to ensure justice and protect fundamental rights in the legal process.

The determination of responsibility, because of the autonomy of artificial intelligence systems, is another issue affecting the application of these systems in decision-making, without a currently resolved consensus. Therefore, a regulatory framework is required to govern the criteria and guarantees for the use of artificial intelligence in the legal domain.

Finally, it is important to note that artificial intelligence is not a threat but an opportunity to enhance the functioning and efficiency of the judicial system, as long as the principles and values governing it are respected. AI can bring benefits such as streamlining processes, optimizing resources, reducing human errors, and improving decision quality. However, it also entails challenges and risks that need to be assessed and prevented through appropriate regulation and human oversight. Artificial intelligence should not replace but complement the role of legal operators, who should be the guardians of justice and fundamental rights.

[52] D. Carneiro, P. Veloso, *Ethics, Transparency, Fairness and the Responsibility of Artificial Intelligence*, in J.F. de Paz Santana, D. Hernández de la Iglesia, A.J. López Rivero (edited by), *New Trends in Disruptive Technologies. Tech Ethics and Artificial Intelligence*, Cham, 2022, 109-120.