Revista Ítalo-Española de Derecho Procesal
pp. 1-18
Madrid, 2026
DOI: 10.37417/rivitsproc/3337
Marcial Pons Ediciones Jurídicas y Sociales

ISSN: 2605-5244
Recibido: 05/02/2026 | Aceptado: 01/04/2026

AI IN JUSTICE: BALANCING EFFICIENCY AND EXPLICABILITY*

Renzo Cavani
Pontificia Universidad Católica del Perú**

ABSTRACT: AI systems intended for use in judicial procedure carry a significant commitment to efficiency—understood from an instrumental perspective—for which accurate predictions are paramount. However, if an AI system has this objective, it is inevitable that explicability will decrease. This means that the program's algorithms are not comprehensible to humans, and often, neither is its output. From an AI ethics standpoint, AI systems applied in the field of justice should possess a significant degree of explicability, yet this compromises efficiency. Therefore, this paper aims to 1) highlight the tension between efficiency and explicability; and 2) argue that, even adopting an AI human-centred approach, opaque systems prioritizing efficiency could be justified, depending on the specific function of the system; specifically, in the greater or lesser degree of interference with the judicial decision that must be justified.

KEYWORDS: Artificial intelligence; judicial process; efficiency; explicability; algorithmic transparency.

---

* This paper is based on the doctoral dissertation entitled "Precedent and Artificial Intelligence: The Impact of Predictive Models on the Theory of Precedent and Judicial Practice", supervised by Professor Jordi Ferrer Beltrán, and submitted to the University of Girona (graded as "excellent *cum laude*"). See Cavani (2025) and also Cavani (in press). An early draft was presented in the IAPL Colloquium 2025 "Judicial Efficiency Revisited", held in Zagreb, Croatia (11-13 September 2025).

** Associate Professor at PUCP. Holds a PhD by Universitat de Girona (Spain) and a LLM degree by Universidade Federal do Rio Grande do Sul (Brazil). Ordinary Member of PRODEJUS-PUCP. CEO and cofounder of Evidence Lab.

1. INTRODUCTION

AI systems in general, and those used in justice in particular, are committed to efficiency, understood from an instrumental perspective (for instance, as a means to an end, but without being qualified as good per se). For this, having the most accurate and precise predictions possible is crucial. However, if this objective is pursued, a reduction in explicability is inevitable: the program's algorithmic functioning is not comprehensible to humans, and often, neither is the output itself.

From an ethical perspective it is said that AI systems applied in justice, given its public relevance, should have a significant degree of transparency to allow for the greatest possible explicability; although this would compromise efficiency.

Therefore, this paper aims to demonstrate two very specific points: 1) the existence of a deep tension between efficiency and explicability (identifying the latter as an end-value of AI, and distinguishing it from other values such as transparency, traceability or accountability); and 2) that, even adopting a human-centred approach to AI, there is justification for AI systems in the realm of justice that prioritize efficiency, sacrificing explicability. However, this will depend on the system's function; specifically, on the greater or lesser degree of interference with the judicial decision that must be justified. Finally, some conclusions are offered.

2. EFFICIENCY AND PREDICTIVE ACCURACY IN AI SYSTEMS

One of the main objectives of current AI systems is to offer increasingly accurate predictions. Generally, these systems are characterized by carrying out machine learning, which involves various techniques, such as decision trees, statistical regression, support vector machines, evolutionary algorithms, neural networks (deep, in the case of large language models, which are a subspecies of generative AI), etc. However, "the activation of these models involves complex calculations, intended to reproduce the correlations between input features and the results to be predicted"; and, while not all these techniques entail the same complexity, "today, the most influential model among those used in machine learning is the neural network. These are computer systems consisting of nodes (so-called neurons) connected by links (also known as parameters) to which numerical weights are assigned (…)"[1].

In other words, the very structure of such systems and the use of big data[2] mean that, as complexity increases, they become more opaque, less com-

[1] Santosuosso & Sartor (2022, p. 1770; 2024, p. 42).

[2] This term refers to "large datasets that are too complex and voluminous to be processed and analysed by traditional data-processing methods" (Khan, 2025, p. 25). Data comes from organizations

prehensible to the human mind, and therefore, less able to clarify their reasoning, whether by explaining or justifying their decisions. These are the so-called black box systems. This is further magnified when the material the models work with consists of unstructured data; that is, data for whose prior processing no human has intervened in classifying the material to be processed by the system[3].

Thus, while there are predictive AI systems in general that offer a fairly clear sequence of steps expressing premises and conclusions (such as the decision tree technique or the rule-based system), the most influential models require more complex calculations to better reproduce the statistical correlations between the input and the result. These calculations are usually expressed in numerical values that determine the system's operation and even serve to train it, as seen when discussing ML. All this leads to the conclusion that if one wanted to institutionally implement an AI system based on statistical frequencies capable of yielding good predictions, the comprehension of that system would necessarily have to be sacrificed.

One of the most important values behind these systems that seek predictive accuracy is *efficiency*. Here, an instrumental sense of the concept of efficiency is assumed; that is, an alignment between the stated goals or objectives and the means to achieve them. There is no intrinsic commitment to the *correctness* of a certain value or objective to be achieved, but rather, with the *best way* to achieve it[4]. So, we can assume that a more accurate prediction generates savings in time, money, and effort, and better results for certain human activities, thereby creating value and ultimately achieving benefits for human beings (beneficence is one of the final values of AI according to various ethical stances). Thus, understood in this way, efficiency would present itself as the best path to well-being; but, by itself, it does not commit to well-being being a correct value: for that, a certain moral philosophy or political theory would be required[5].

(costumer information, transactions), humans (social networks, media, photos, internet searches, e-mail, etc) and, above all, machines (sensors, video cameras video, satellites, etc) (Gupta & Mamta, 2024, pp. 3 ss.). Data can arrive in various forms: structured, semi-structured, or unstructured. The latter two forms are the most prevalent, and their processing requires highly specialized techniques (see EMC2, 2015, p. 5). This led to the creation of a multidisciplinary field aimed at using big data to extract information and influence decisions: big data analytics. Several techniques were developed in this area, the most common of which include data mining, machine learning, natural language processing, data visualization, predictive analytics, statistical analysis, and real-time analytics, among others (see Demirbaga *et al.*, 2024, pp. 3-4). Therefore, *big data analytics* and IA are closely connected.

[3] Gioia (2025) argues that *big data* would be a "condition of possibility for the use of IA in judicial system" (p. 1354), but this is not entirely accurate. Indeed, if the goal is to achieve a high degree of efficiency, more sophisticated programs requiring the processing of large amounts of data will be necessary. Nevertheless, there are systems that do not necessarily rely on big data and can still provide satisfactory results, as exist and are applied in several judicial systems.

[4] See Giabardo (2022).

[5] See Giabardo (2022, pp. 57-58).

3. THE TENSION BETWEEN EFFICIENCY AND EXPLICABILITY

In the framework of developing AI systems, at a general level, it is often stated that there is a tension between two objectives to be pursued: a) providing more accurate predictions (for instance, efficiency, as we have just seen); and b) providing better explanations for said predictions. As mentioned, if one wishes to prioritize the first objective, the second will necessarily be sacrificed[6]; and, conversely, if the aim is to have systems that can provide more and better explanations, the accuracy of the predictions could end up being sacrificed, thereby compromising efficiency. For its part, choosing *how opaque* —or, also, *how transparent*—a system should be involves various ethical considerations. It is certainly not easy to strike a balance between efficiency and explicability[7].

Explicability is seen as a core principle of AI ethics[8]. Luciano Floridi, from a bioethics perspective[9], proposes that the central ethical principles of AI (which are often heralded in various declarations of principles[10]) be beneficence, non-maleficence, autonomy, and justice, to which, precisely, explicability is added[11]. According to the author,

> explicability is understood as incorporating both the *epistemological* sense of *intelligibility* —as an answer to the question "how does it work?"—and in the

[6] Santosuosso and Sartor (2022, pp. 1770 ss.).

[7] Sartor and Santosuosso (2022, p. 1771; 2024, p. 47). A similar, although not identical, tension is argued by Floridi (2025). The author asserts that there is a trade-off between certainty or accuracy and scope, in the sense that no AI system can simultaneously maximize epistemic certainty (i.e., error reduction) and the scope of the system's mapping. Thus, he considers that while symbolic AI has high certainty but low scope, generative AI has high scope but low certainty. This leads him to conclude that generative AI hallucinations are not "errors" of the system, but rather an inherent feature resulting from having such a broad scope. Therefore, seeking an "error-free" system of this kind is impossible, which is why a hybrid system would be recommended.

For her part, Smuha (2025) recently has presented a tension between the efficiency or optimization sought by AI systems used by state entities and fundamental rights, whose protection is a duty of the very State that employs these systems. Thus, according to the author, algorithmic regulation can erode legality (opacity of norms), hinder the citizen's ability to understand and challenge decisions, and affect legitimacy, in the sense that the moral authority of the law is perverted, turning it into an efficient yet arbitrary tool of social control (here lies the wordplay: the law viewed in this way would no longer be a *Rule of Law*, but a *Rule By Law*). Hence, according to Smuha, it is necessary to set strict limits to continue preserving freedom and restraining state power. See also, Smuha (in press). However, in neither of those important contributions the author stresses any distinction between AI systems according to its functions. This is important because, as we will see, the impact is not the same: the risk of a "Rule by Law" might be overcome if we solve correctly the tension between efficiency and explicability, as I will try to show.

[8] Analyzing several opposite views on this point, see Buijsman, Klenk & Van Den Hoven (2025, p. 63).

[9] Floridi cites Beauchamp and Childress (2013). See also the later edition: Beauchamp and Childress (2019).

[10] See Asilomar principles (Future of Life Institute, 2017), Montreal Declaration (Université de Montreal, 2017), among others.

[11] Floridi (2023, pp. 57 ss.)

ethical sense of *accountability*—as an answer to the question "who is responsible for the way it works?" The extra principle is required by the fact that AI is a *new* form of agency[12].

The goal of achieving explicability can be approached from two different perspectives: the explicability of the system's functioning and, on the other hand, the explicability of the output[13]. On one hand, there is the algorithmic functioning, which is not only limited to indicating the type of inputs fed to the system, but also the statistical and mathematical models effectively employed. While achieving this is possible in low-complexity systems (such as those based on decision trees), it is no longer the case in high-complexity systems, like those being implemented in recent years, many of them based on unsupervised machine learning or on neural networks and other complex statistical techniques, with results that not even the designer can understand or explain. And all this is more complex considering those systems whose training is protected by trade secrets, as is the case of ChatGPT, that belongs to the company OpenAI[14].

Thus, faced with a reality in which AI systems necessarily have a high degree of opacity, various efforts have been made seeking greater understanding and transparency—and, hence, explicability—of the system. This is the field of Explainable AI or "XAI"[15]. The importance of systems providing an explanation or interpretation of their results in a way that generates user understanding is problematized by XAI[16]; but it is not about the systems themselves offering explanations of how that result was reached, nor the reasons for the system's decision. It is the human developers of the system themselves

[12] Floridi (2023, p. 57). For a taxonomy of ethical principles (with special attention to AI systems in the sphere of justice), see Simón Castellano (2023, pp. 179 ss.), although the author refers to them, not without some imprecision, as "legal guarantees." This point is further developed in Simón Castellano (2023a).

On the other hand, it is possible to distinguish the concept of an agent (agency) discussed in the philosophy of mind and in psychology, from a more minimalist concept used in computer science. In this field, according to Floridi (2023) " [it] requires that a system satisfies only three basic conditions: it can a) receive and use data from the environment, through sensors or other forms of data input; b) take actions based on the input data, autonomously, to achieve goals, through actuators or other forms of output, and c) improve its performance by learning from its interactions" (p. 10).

Therefore, it is possible to speak of *biological* agents (a human or a dog), *social* agents (a company or a government), and *artificial* agents (a bot). Thus, artificial intelligence (in the weak sense) is not about creating or reproducing human intelligence, but rather about generating results *without it*, in an ever-increasing number of activities. AI would thus be a *new form of agent, but not a new form of intelligence* (p. 20 ss.). The impact of this new agent in the society, and its 'normative power' precisely justifies the shaping of ethical principles that regulate that impact. See Lupo (2022).

[13] Páez (in press).

[14] Páez (in press).

[15] Strictly speaking, XAI is a branch of computer science that aims to make opaque models —specifically their outputs—more explainable, interpretable, and transparent for users. The literature on this topic has been growing considerably: See Gunning *et al.* (2019); Páez (2019, pp. 444 ss.; Paéz, in press); Barredo Arrieta *et al.* (2020); Casacuberta *et al.* (2024); Buijsman (2022); Boge & Mosig (2025).

[16] The various types of XAI models are critically addressed by Páez (2019). Meanwhile, Meert *et al.* (2025, pp. 37 ss.) discuss AI systems designed to bring explicability to otherwise opaque AI systems.

who must explain (or, if appropriate, justify) to other humans (the users of the technology) what they do and have decided to do with the AI[17].

One way to "make explicit" is for the reasons to be contained in statements that are part of the system's output, so that they allow for a proper understanding of the reasoning carried out by the system. Take, for example, of AI systems that work with precedents, serving as a guide for judges to make decisions[18]. Here it is not enough for the statement to simply say "The case should be decided in favour of the plaintiff; to do so, follow precedent X". It must indicate *to what extent* precedent X justifies deciding the current case in favour of the plaintiff, just as a human judge would be expected to do.

It is often said that explicability includes other concepts such as transparency, interpretability, intelligibility, or even accountability. It would ultimately be another way of expressing all of them[19]. However, it is possible to distinguish at least two different problems surrounding explicability: a) the first consisting of accounting for the functioning of a system, which includes offering enough information to users, accountability mechanisms, ensuring there is no algorithmic discrimination, etc.; and b) the second, regarding the possibility of understanding the system's functioning, the inputs entered, and the outputs obtained (which, of course, will be a matter of degree[20]). It could be said, then, that explaining the system's functioning, understanding it as much as possible, and interpreting it in a sound manner *contributes* ultimately to achieve explicability. From this perspective, transparency, intelligibility, and accountability might be understood as instrumental values that aim for explicability, with the latter being understood as an end-value. However, a conceptual precision is mandatory.

---

[17] As Coeckelbergh (2020) stressed: "Yet if the goal is not to have *machines* explain, but rather demand this from *human beings* who are able to explain things to other human beings, then there is a chance that explainable AI may work. Responsibility then does not mean that one explains *everything* that contributed to the action or decision, but rather that one can know and select what is relevant to what the other (human being) wants and needs to know. In the case of AI and other advanced automation systems, this *can* be done by humans if and only if (a) those humans are sufficiently supported by technical systems that are transparent *enough* for the (primary) purpose of humans explaining things to other humans (there is no demand for absolute transparency here), and (b) those humans are sufficiently willing, capable, and educated to imagine and understand what those affected by the technology may ask and demand from them. This can be supported by actually asking stakeholders what kind of explanations they actually want and need. The assumption is then that only humans can really explain and *should* explain what they decide and do, and that explanation itself is deeply social and relational (p. 2054; the italics are from the author)".

The author later notes that, in addition to explanations, users would also need *reasons*, and that, in terms of the comprehension requirement, these would function in the same way as explanations.

[18] These are systems that have been under development since the late 1980s. For an overview, see Bench-Capon (2017); Bench-Capon *et al.* (2024). For the most recent studies, see Gray, Šavelka, Oliver & Ashley (2022, 2023, 2023a, 2024, 2024a); Gray, Li and Ashley (2025); Li, Gray, Šavelka and Ashley (2025). These models were the focus of my doctoral thesis, defended at the University of Girona: Cavani (2025). See also Cavani (in press).

[19] See Floridi (2023, p. 63). In later works the author changed its conceptual view, as we will see later.

[20] Páez (in press).

In the case of transparency, understood as the "accessibility of information about an IA system's components and architecture"[21] at the level of broader discussions on information ethics, it is seen as a condition for achieving other values such as accountability, security, well-being, or informed consent[22]. In fact, in the field of AI, transparency is a typical case of a condition of possibility for other ethical principles; however, complete and total transparency could also be problematic: while it would give users critical information about the system's functioning, it could also overwhelm them, thereby generating an unintentional opacity of the system[23], and even increase costs, or allow the system to be tricked, generating inequities[24]. Therefore, if on one hand explicability is an end-value, seen from this perspective, transparency could not be. In this view, a certain *degree* of transparency is a necessary, but not sufficient, condition for explicability[25].

There are some AI systems that not only allow the user to understand the why of the output's meaning but also offer *justifications*; that is, reasons or guides employed to reach a certain decision. However, as previously noted, it is important to bear in mind that not all explicability is directed at the end-user, nor is it reduced to offering only justifications.

In fact, explicability can be directed at certain operators who can technically understand the system's functioning, and is not limited merely to the output itself[26]. On the other hand, due to the system's particularities, there will not always be justifications but rather genuine predictions. In fact, not every output yields a decision based on reasons[27]; it can also provide information so that the recipient can make a decision. This is the case with legal assistants such as SCOTUS or Lex Machina which, by working with periphe-

[21] Buttaboni & Floridi (2026, p. 3). For example, the Asilomar AI Principles, under the heading of Ethics and Values, expressly include judicial transparency: "Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority". However, here is a conceptual confusion between transparency and explicability, as it is common in other legislations, both soft and hard law, and in the scholar literature.

[22] Turilli and Floridi (2009, p. 107). More deeply, see Floridi (2012, 2013).

[23] Floridi (2023, pp. 98-99).

[24] Floridi (2023, p. 99).

[25] For Buttaboni & Floridi (2026), transparency and traceability must be distinguished: whilst the former means the access to information about the system, the latter presupposes the former, in order to *understand*, from a causal perspective, the system's processes. According to the authors, "traceability presupposes a degree of transparency —namely, access to relevant training data and system architecture—but introduces a temporal and procedural dimension: the capacity to track and reconstruct what the system did, when, and why. It transforms static access into dynamic reconstruction, making it indispensable for auditability and effective risk management" (p. 5).

[26] This is what Buttaboni & Floridi (2026) name *interpretability* (p. 6).

[27] As it seems to be understood by Buttaboni & Floridi (2026), to whom explicability is the capacity of justifying an specific outcome to an affected user, allowing her to control and question that assessment or decision. Explicability is, thus, a teleological concept. According to the authors, "a system becomes explainable only when it can provide justifications at the appropriate level of abstraction for a specific user's needs and capabilities"; "AIS [artificial intelligence system] is explainable when a system provider (e.g., the model developer) generates a justification of the output that provide epistemically adequate grounds for the recipient (e.g., the client, regulator, or affected party) to assess, act upon, and crucially, challenge that given output" (p. 7).

ral elements of a case (for example: the name of the judge, the parties, the law firms involved, the judicial district where the claim was filed, whether there was a settlement and its value), offer information so that a lawyer can strategically decide the course of their client's future case.

In any case, values such as transparency, traceability, or the understanding of algorithmic processes by developers do not necessarily lead to explicability on the part of the user; that is, they are not sufficient conditions. What is required is to promote an understanding of the system that allows the user to challenge, in a relevant way, the decision that ultimately affects them[28].

It is difficult to achieve a general ethical theory for all systems that use AI. As we have seen, there will be some AI systems (at a general level, not restricted to legal field) that, due to their own algorithmic design characteristics, will intrinsically generate problems of a lack of transparency and explicability. The influencing factors are the human cognitive impossibility of interpreting algorithmic models and databases; the lack of tools to visualize and understand codes and data; data whose structure does not allow them to be read; updates to the models and human influence; the fact that the lack of transparency is intrinsic to the system given its self-learning process; and the malleability of algorithms due to their reprogramming, which is fundamental for the improvement of their results, among others[29]. And if it is possible to know and describe the algorithmic design, considering the technical and complexity level with which they are designed, it is also a matter of deciding whether the public whose legal sphere will be affected by the AI should know and understand the algorithmic design[30]. If the answer is affirmative, the additional problem of the *degree of knowledge* to be promoted among the users arises; that is, whether to promote a broad knowledge of the design, development, and implementation of the algorithms, or rather a more limited knowledge (always based on the premise that transparency, at the level of AI ethics, is not an end-value to be achieved in itself[31]). Also, it will be influential whether such programs are public or private (for the purposes of commercial or industrial secrets) and what the specific use of said program would be. This point is particularly relevant for what I will argue later.

---

[28] According to Buttaboni & Floridi (2026), this presupposes an epistemic dimension of explicability (pp. 7, 10).

[29] Floridi (2023, pp. 97-98).

[30] Gina Gioia's proposal seems somewhat excessive in asserting that the only way to preserve judicial independence is for the judge to exercise "constant control over all phases of the process: from data collection, to algorithm programming, up to operational management" (2024, p. 1354). The judge is neither a technician nor a software engineer and, as I have been arguing, not even someone with deep technical knowledge could comprehend all the phases of an algorithmic process.

[31] From a broader perspective, Floridi (2023, pp. 94-95) identifies six ethical concerns in the design, development, and implementation of AI algorithms. They can be grouped into two domains: epistemic factors and normative factors. The first includes problems related to inconclusive evidence, inscrutable evidence, and misguided evidence; the second includes unfair outcomes and unintended consequences. These five factors lead to hindering "the possibility of identifying a cause for an outcome and thus attributing moral responsibility for it" (p. 95). This specific issue constitutes the sixth concern: traceability. According to Floridi, opacity or lack of transparency is linked to both normative factors.

But alongside transparency, it is possible to identify another value linked to explicability: *controllability*. This is particularly relevant when we talk about AI systems applied in the field of justice. Controllability requires that the reasons for a certain AI-based decision or action must be understandable so that they, in turn, can be challenged by the users or citizens who will suffer their effects. Consider the case of a judicial decision. If the judge adopts (or rejects) as a premise the reasons offered by the system about the correct outcome of the ruling based on comparisons with past cases, then this reason must also be subject to control. Although, it is not only a control by the general public given the public nature of judicial decisions (which could be resolved through periodic audits by the programmers), but also the control through the appeal mechanisms made available to the parties who suffer the consequences of the said decision. Therefore, controllability is also an instrumental value: the understanding and control of judicial decisions serve, in turn, to make decisions regarding the calculability of the responses of the judges who use the system, so that the affected party can achieve a lawful decision through the appeal[32].

4. RESOLVING THE TENSION? THE CRITERION OF THE SYSTEM'S FUNCTION

When it comes to public decisions, as Santosuosso and Sartor indicate, it is generally preferable "a system that commits fewer errors versus a more fallible system that is capable of providing reasons for its choices". However, as the authors continue,

> The requirement to provide explanations is, however, fundamental when the system is used for the exercise of functions of public relevance, especially when conflicting interests are at stake and the needs for control are preeminent, as is the case with administrative action and even more so in the sphere of jurisdiction[33].

It is true that AI systems in the field of justice are inserted into a public sphere and, therefore, to a greater or lesser extent, have a great impact on citizens. However, to resolve the tension between explicability and efficiency, it seems to me that the criterion of whether the system is (or not) used to exercise a public function is not sufficient. The reason for this is that it does not tell us *how explainable* the system should be or, which is the same, *what degree of opacity* we should be willing to tolerate for efficiency to prevail. Likewise, it does not seem to distinguish between systems applied in bureaucratic procedures of justice system management and those used in the judicial process; and, in this context, there is no distinction between systems applied to procedural or bureaucratic aspects and those that affect the decisional sphere, that is, the judge's own decision-making processes.

---

[32] This is problematized by Bujosa Vadell (2022, pp. 751 ss.).

[33] Santosuosso & Sartor (2022, p. 1771). See also Santosuosso & Sartor (2024, p. 47). Later on, the authors clarify better their argument concerning the tension between explicability and efficiency.

Take, for example, AI tools that allow building statistics and predictions from the measurement of data collected from various courts and tribunals, or instruments that contributes to the selection of personnel to be hired. On the other hand, consider systems that help manage cases, that summarize hearings, that draft judicial resolutions, those that propose certain measures based on a risk assessment, and also legal assistants that offer reasons for deciding. All of them have a more direct or indirect influence on judicial decisions that, in the end, will ascertain the legal situation of the parties and guide social behaviour.

I think it is of little use to argue, without further ado, that an AI system must be transparent or explainable if it is to be used in the field of justice. Actually, a more precise question is the following: *How transparent/explainable should that system be?* Or, also: *To what extent should efficiency be prioritized over the possibility of achieving greater explicability?* The reason for this is that, in general, there is an underlying purpose regarding efficiency in any AI system, from a general point of view, and also in any AI system applied to the field of justice. No AI system is designed or implemented to increase the costs or time dedicated to producing the expected result, nor to generate greater effort, for example, by having to review or control everything the system does. If a system has any of these effects, it should be discarded.

Nevertheless, this does not mean that, because every AI system has a purpose oriented towards efficiency, the tension between it and explicability must be resolved in favour of the former. Like intelligibility or understanding regarding the system's functioning, *the achievement of efficiency is also gradual*. It is possible, therefore, that this tension might be oriented towards explicability, restraining efficiency, but without completely renouncing to it. [34] It is also possible to give up explicability to a certain extent, with a greater margin of opacity, in order to favour efficiency; but, in the same way, without discarding any type of explanation, including the possibility of a general description of the system's functioning [35].

But, what would be the criterion for opting for one alternative or the other? My proposal is to assume the criterion of the system's function (that is: what is it for?). In general, it is about *ascertain the degree of interference* that the system's results have on the judge's decisions throughout the process (and not just the final ruling). Thus, the lower the interference, the more efficiency can prevail; in contrast, the more direct the interference, the greater the explicability that should be promoted. This lies not necessarily in any risk prevention according to the harm the system could cause (although this

[34] Conversely, Gioia (2024, p. 1356) argues: "every system intended to assist judicial activity must have a clear and transparent origin, structure, and functioning, and be approved by law in a specific rather than generic manner". Here, the author makes no distinction between systems with greater or lesser interference and, in my view, neither does she base her conclusion on the European Regulation, which would indeed allow for such a differentiation in terms of high-risk or acceptable-risk systems, as we will see later.

[35] Santosuosso & Sartor (2024, pp. 132-133).

could exist); in fact, it has more to do with the duty to provide reasons, that is, to justify decisions. This is an activity entrusted to the judge due to a series of political agreements on the organization of the system and legal and social rules[36]. If we have an AI system that facilitates or contributes with information, predictions, or reasons so that the judge can fulfil their job of deciding, it becomes necessary for the public and the parties to know enough about said system. Note that this is even more necessary when an AI system is capable (partially or totally) automating some reasoning processes that the judge executes to shape the justification of their decision, given that the judge could confer some *auctoritas* to the machine or develop a passive submission to it[37].

Once this is defined, it would be appropriate to ascertain the degree of explicability that should be achieved: whether of the algorithmic functioning itself or of the output; whether a general or an individualized description is sufficient; whether it is about an explicability focused on the user of the system or if it focuses more specifically on the parties, etc. Likewise, two subcriteria for deciding on the degree of explicability could be not to excessively compromise the system's effectiveness nor generate unsustainable costs; that is, that they are "compatible with the available resources and with commercialization requirements"[38]

An example of systems that are far from conditioning the direction of a decision or being, in any way, decisive in the decision-making processes, are those intended to improve bureaucratic procedures of the judiciary, to manage and classify files according to efficiency criteria, distribute workload, transcribe hearings, etc[39].

Now take as an example an AI system that reasons based on precedents, offering the judge reasons to decide new cases based on factors found in previous cases[40]. Here, undoubtedly, the degree of the system's explicability will depend on the complexity of the technique used, the amount of data involved, and the desirable knowledge depending on the target audience (namely, judges, system users, lawyers, and litigants). This is even more complex given that legal knowledge and reasoning itself can become so sophisticated that it is not something that can be easily grasped by subjects who are not legal operators.

---

[36] Santosuosso & Sartor (2024, p. 143).

[37] Santosuosso & Sartor (2024, p. 143).

[38] Santosuosso & Sartor (2024, p. 136). The authors also understand the duty of explicability as a *best-effort* duty.

[39] The European AI Regulation is in this same vein: "the final decision-making must remain a human-driven activity. The classification of AI systems as high-risk should not, however, extend to AI systems intended for purely ancillary administrative activities that do not affect the actual administration of justice in individual cases, such as anonymisation or pseudonymisation of judicial decisions, documents or data, communication between personnel, administrative tasks" (paragraph 61).

This is also remarked by Ariza Colmenarejo (2023, p. 34). Thus, contrary to what Nieva Fenoll stresses (2025, p. 13), in this point it is not mandatory to have a "white box", because of the diminishing of the system's efficiency.

[40] As previously remarked, this was my main concern in Cavani (2025). See also Cavani (in press).

Notwithstanding this, if we are faced with this type of AI system, committed to providing reasons for a judge to decide, it is necessary to provide sufficient information about the algorithmic functioning of the system as part of the public policy for using said model. In my opinion, in terms of explicability, the sufficiency of this knowledge requires that it be rather broad[41]. It must be explained how the algorithms have been trained; under what criteria the dataset has been built; which cases compose it; what factors have been designed to represent the cases; the computer science techniques that were used must be detailed; and what type of statements the system is capable of offering as output. Likewise, it must be made transparent what the user judge will do, that is, what type of input she will enter into the system; how she will be able to control the reasoning operations executed by the system; and, above all, how she would incorporate its reasons into her decisions. For this, of course, many protocols and regulations designed primarily for the litigants will be required[42].

Concerning this requirement of sufficient explicability, AI-based assistants that, although capable of offering much more precise and refined reasons, are fundamentally opaque and do not allow a reasonable degree of understanding, should be excluded. In fact, just as it is beneficial that more sophisticated techniques be implemented for the algorithm training and the handling of data (for example: standardizing the language of decisions for an automated representation of factors), it is also necessary to develop techniques that point towards a high degree of explicability[43].

Thus, faced with an eventual implementation of an AI system that serves as a judge's legal assistant, a necessary requirement to achieve the greatest possible transparency and controllability is that the explanation of reasons be

[41] According to Santosuosso & Sartor (2024): "A prediction unaccompanied by an explanation may perhaps be sufficient for a party interested in establishing whether it is worth suing or defending in court, but not for the judge, who has the duty to decide based on the facts and the law, and to justify their decision. However, even for the party, relying on an oracular judgment without being able to know and evaluate the reasons supporting one alternative or the other can be quite reckless. At most, it can be an input for further consideration. On the other hand, a proposed outcome accompanied by reasons, and indeed the identification of multiple alternative outcomes —each accompanied by an assessment of its plausibility, probability, and coherence with past decisions—can be useful to both the parties and the judge" (p. 130).

[42] This is not to say that knowledge regarding the system's technical functioning—that is, the kind understood by specialized developers—should be of secondary importance. This information can readily be presented in a manual or guide designed for non-specialists in computer science, even highlighting certain areas of the system's opacity.

[43] Although this is currently very difficult, it is not impossible (Floridi 2023, pp. 100 ss.). However, it must always be considered that "since an ML algorithm learns rules based on statistical regularities that may surprise humans, its rules may not necessarily seem reasonable to humans. ML predictions are data-driven. Sometimes the data contain features that, for spurious reasons such as coincidence or biased selection, happen to be associated with the outcomes of cases in a particular collection. Although the machine-induced rules may lead to accurate predictions, they do not refer to human expertise and may not be as intelligible to humans as an expert's manually constructed rules. Since the rules the ML algorithm infers do not necessarily reflect explicit legal knowledge or expertise, they may not correspond to a human expert's criteria of reasonableness". (Ashley, 2017, p. 111).

made by the system itself[44]. At this point an issue of efficiency comes into play. It would not be enough for these justifying reasons to be freely reconstructed by each of the users, as this would lose the real utility of the computerized assistant. The reasons must be contained in the system's own statements, be incorporated (or not) as a premise of the judicial decision when the judge constructs the reasoning, motivating why they follow (or not) the system, and must be made known to the parties. Thus, without prejudice to the knowledge of the system's algorithmic functioning, it is necessary to know how the system has reasoned in each specific case, being able to access, literally, the statements produced as output. This will allow for control over, also: a) the type of input introduced by the judge to activate the system; b) the database with which the system works; and c) the type of reasoning operations carried out by the system for the specific case (specifically: the comparisons and distinctions made between the current case and the preceding cases).

Finally, a different problem will be the degree of sufficiency of the explication of reasons in the output. The purpose is to generate a greater understanding for the user who makes the decision (that is, the judge), the legal operator or legal expert, the citizens whose cases will be resolved based on decisional premises placed by this system, and also, the general public who will evaluate the use and functioning of the system.

5. CONCLUSIONS

Even if a human-centred AI approach is assumed and we agree on certain core principles or values that should guide the design and use of AI systems, the demands for efficiency and explicability in the context of justice must depend on the program's function. In this way, systems with more accurate predictions that bring greater efficiency could be justified, even if there is some sacrifice of transparency and explicability, as could be the case with systems that organize and classify processes according to certain characteristics or that prepare decision templates. As has been seen, this would not apply to AI systems that have a more direct impact on the decision because the need for justification requires that the functioning of these systems be explained to a sufficient degree to understand it and be able to perform control actions.

It is true that so far there is nothing similar to an "automated decision" or a "robotic decision" that implies a total replacement of the judge by AI[45]; but

[44] This proposal (which cannot be fully developed here) is in line with the Montreal Declaration for a Responsible Development of AI (Université de Montréal, 2018), specifically its principle of democratic participation.

[45] For instance, Luís Greco (2020) argues that the "robot judge" is factually feasible but ethically indefensible. The problem with this perspective is that the very concept of a "robot judge" remains unclear. Are we referring to a machine capable of completely replacing the human judge? If this is the premise, it seems highly unlikely to be achieved given how AI has been structured thus far. To give just one example: an AI capable of assessing evidence as we do would need to possess the capacity of repre-

some reasoning processes can be verified that could be carried out entirely by the machine, even if they are under the supervision of the judge. Precisely for this reason, it could be said that there is not yet a genuine revolution in legal practice; although this does not mean that technology is not advancing and, eventually, that AI systems powerful enough to transform the way judicial decisions are justified will be developed. There is no doubt that this will increase the tension between explicability and efficiency (which, in this hypothesis, will be enhanced). Faced with this, in my opinion, it is very important to be clear about the end-values that AI as a discipline should pursue and, regardless of the complexity of the computerized model to be implemented, to try to resolve the tension based on the interference it has in the making of the judicial decision.

Last but not least, the criterion proposed here—that is, resolving the tension based on the system's degree of interference with the judicial decision—broadly aligns with the risk-based approach adopted by the European Artificial Intelligence Act (AIA)[46]. While the AIA generally classifies AI systems intended to assist judicial authorities in interpreting facts and the law as "high-risk" (Annex III, point 8), Article 6(3) introduces an exception for systems that do not "substantially influence the outcome of decision making" (such as those performing narrow procedural or preparatory tasks). Thus, under the AIA framework, AI tools with a lower degree of interference in the judge's reasoning would pose a more limited risk, justifying a greater allowance for efficiency over strict explicability, whereas systems that directly shape the substantive decision would demand maximum transparency and human oversight[47].

6. REFERENCES

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Era.* Cambridge University Press.

Ariza Colmenarejo, M. J. (2023). Hacia la eficiencia digital y el refuerzo de las garantías procesales. In J. Caro Catalán, A. Sánchez Rubio, C. Durán Silva, & P. Pulido Manuz (Eds.), *La optimización de la justicia: Reflexiones sobre la eficiencia procesal* (pp. 15-36). Fundación Privada Manuel Serra Domínguez-Atelier.

senting the world in order to apply rules of experience, and be designed not merely to find correlations, but also to establish causalities and formulate inferences.

Regarding the latter Greco's argument, it overlooks several important aspects. Indeed, if we are not dealing with a truly intelligent agent (and, therefore, one lacking both authorship over the decision and responsibility for it), it would be a mistake to accept a complete replacement as appropriate. But what about procedural or simple interlocutory decisions? For instance, could there be a replacement coupled with human oversight? Ultimately, I believe the real ethical issues lie not so much in a complete substitution, but rather in "partial substitution"—that is, in AI systems functioning as decision-making assistants and, therefore, capable of replacing certain reasoning processes that serve as premises for constructing the final decision.

[46] For a comprehensive analysis of the AIA's scope and impact assessment, see Van Dijck (2022); Mökander *et al.* (2022); Floridi (2023, pp. 85 ss.); Laux (2023); Novelli *et al.* (2024); Walters *et al.* (2024); Hupont *et al.* (2024); Gatt *et al.* (2025); Finck (2026).

[47] I will try to delve more on this argument in a future work.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Beauchamp, T. L., & Childress, J. F. (2012). *Principles of Biomedical Ethics* (7th ed.). Oxford University Press.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of Biomedical Ethics* (8th ed.). Oxford University Press.

Bench-Capon, T. J. M. (2017). HYPO'S Legacy: Introduction to the Virtual Special Issue. *Artificial Intelligence and Law, 25,* 205-250. https://doi.org/10.1007/s10506-017-9201-1

Bench-Capon, T. J. M., Atkinson, K., Bex, F., Prakken, H., & Verheij, H. B. (2024). Computational models of legal argument. In D. Gabbay, G. Kern-Isberner, G. R. Simari & M. Thimm (Eds.). *Handbook of Formal Argumentation*, Vol. 3 (pp. 101-123). College Publications.

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines, 32,* 563-584. https://doi.org/10.1007/s11023-022-09607-9

Buijsman, S., Klenk, M., & Van den Hoven, J. (2025). Ethics of AI: Toward a "design for values" approach. In N. Smuha (Ed.), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence* (pp. 59-78). Cambridge University Press.

Bujosa Vadell, L. (2022). Ética e inteligencia artificial: una mirada desde el proceso jurisdiccional. *Revista eletrônica de direito processual, 23*(1), 733-768. https://doi.org/10.12957/redp.2022.64391

Buttaboni, C. & Floridi, L. (2026). A regulatory taxonomy of AI opacity in the EU: Rethinking transparency, traceability, interpretability, and explainability. *AI and Ethics, 6*, 100. https://doi.org/10.1007/s43681-025-00940-0

Casacuberta, D., Guersenzvaig, A., & Moyano-Fernández, C. (2024). Justificatory explanations in machine learning: For increased transparency through documenting how key concepts drive and underpin design and engineering decisions. *AI & Society*, *39*, 279-293. https://doi.org/10.1007/s00146-022-01389-z

Cavani, R. (2025). *Precedente e inteligencia artificial: El impacto de los modelos predictivos en la teoría del precedente y la práctica judicial* [Doctoral dissertation, *Universitat de Girona*]. https://www.tdx.cat/handle/10803/694941

Cavani, R. (forthcoming). *Precedentes e inteligencia artificial: de la predicción a la justificación.* Marcial Pons.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*, 2051-2068. https://doi.org/10.1007/s11948-019-00146-8

Demirbaga, Ü., Sungh Aujla, G., Jindal, A., & Kalyon, O. (2024). *Big Data Analytics. Theory, Techniques, Platforms, and Applications.* Springer.

EMC2 (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* EMC Education Services-Wiley.

Finck, M. (2026): *The EU Artificial Intelligence Act: A Commentary.* Oxford University Press.

Floridi, L. (ed.) (2012). *The Cambridge Handbook of Information and Computer Ethics.* Cambridge University Press.

Floridi, L. (2013). *The Ethics of Information.* Oxford University Press.

Floridi, L. (2023). *The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities*. Oxford University Press.

Future of Life Institute (2017). *Asilomar AI Principles* www.futureoflife.org/open-letter/ai-principles/

Gatt, L., Lo Conte, M. T., & Mazzarella, M. E. (2025). Analysis and perspectives on the AI Act: scope and impact assessment. *European Journal of Privacy Law & Technology*, 2, 1-39.

Giabardo, C. V. (2022). Efficiency as a means and as and end: Some thoughts on happiness, wealth and justice in the legal economic debate. In D. Almanza Torres & M. Pereira Ribeiro (Coords.). *A economia comportamental aplicada ao direito. Estudos sobre racionalidade, políticas públicas e leis* (pp. 57-74). Ithala.

Gioia, G. (2025). Intelligenza artificiale e indipendenza della magistratura: un equilibrio tra trasparenza e controllo. *Rivista di Diritto Processuale*, 80, 4, 1338-1363.

Gray, M., Šavelka, J., Oliver, W., & Ashley, K. D. (2022). Toward automatically identifying legally relevant factors. In E. Francesconi, G. Borges & C. Sorge (Eds.), *Legal Knowledge and Information Systems* (pp. 53-62). IOS Press. https://doi.org/10.3233/FAIA220448

Gray, M., Šavelka, J., Oliver, W., & Ashley, K. D. (2023). Automatic identification and empirical analysis of legally relevant factors. In *ICAIL '23: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (pp. 101-110). ACM. https://doi.org/10.1145/3594536.3595157

Gray, M., Šavelka, J., Oliver, W., & Ashley, K. D. (2023a). Can GPT alleviate the burden of annotation? In G. Sileno, G. Spanakis & G. van Dijck (Eds.), *Legal Knowledge and Information Systems* (pp. 157-166). IOS Press. https://doi.org/10.3233/FAIA230961

Gray, M., Šavelka, J., Oliver, W., & Ashley, K. D. (2024). Empirical legal analysis simplified: Reducing complexity through automatic identification and evaluation of legally relevant factors, *Philosophical Transactions of the Royal Society A. Mathematical, Physical, and Engineering Sciences, 382*(2270), 20230155. https://doi.org/10.1098/rsta.2023.0155

Gray, M., Šavelka, J., Oliver, W., & Ashley, K. D. (2024a). Using LLMs to discover legal factors, *arXiv:*2410.07504v1.

Gray, M., Li, Z., & Ashley, K. D. (2025). Generating case-based legal arguments with LLMs. In *CSLAW'25: Proceedings of the 2025 Symposium on Computer Science and Law, March 25–27, 2025, München, Germany* (pp. 160-168). ACM. https://doi.org/10.1145/3709025.3712216

Greco, L. (2020). *Poder de julgar sem responsabilidade do julgador: a impossibilidade jurídica do juiz-robô.* Marcial Pons.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G-Z (2019). XAI - Explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Gupta, B. B., & Mamta (2024). *Big data: Management and Analytics.* World Scientific.

Laux, J. (2023). Institutionalised Distrust and Human Oversight of Artificial Intelligence: Towards a Democratic Design of AI Governance Under the European Union AI Act. *AI & Society*, 39, 2853–2866 https://doi.org/10.1007/s00146-023-01777-z

Li, Z., Gray, M., Šavelka, J., & Ashley, K. D. (2025). Measuring faithfulness and abstention: An automated pipeline for evaluating LLM-generated 3-ply case-based legal arguments. *arXiv:2506.00694v2*.

Lupo, G. (2022). The ethics of artificial intelligence: An analysis of ethical frameworks disciplining AI in justice and other contexts of application. *Oñati Socio-Legal Series, 12*(3), 614-653. https://doi.org/10.35295/osls.iisl/0000-0000-0000-1273

Hupont, I.; Fernández-Lorca, D.; Baldassari, S. & Gómez, E. (2024). Use Case Cards: A Use Case Reporting Framework Inspired by the European AI Act. *Ethics and Information Technology*, 26, 19. https://doi.org/10.1007/s10676-024-09757-7

Meert, W., De Laet, T., & De Raedt, L. (2025). Artificial intelligence: A perspective from the field. In N. A. Smuha (Ed.). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence* (pp. 17-39). Cambridge University Press.

Mökander, J.; Juneja, P.; Walton, D. S. & Floridi, L. (2022). The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What Can They Learn from Each Other? *Minds and Machines*, 32, 751-758. https://doi.org/10.1007/s11023-022-09612-y

Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M. & Floridi, L. (2024). AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. *Digital Society*, 3, 13 (*original paper*). https://doi.org/10.1007/s00146-023-01723-z

Nieva Fenoll, J. (2025). Inteligencia artificial generativa (IAGen) y defensa: ¿Cómo lograr la "convicción" de una máquina? *Revista Italo-Española de Derecho Procesal*, 1-17. https://doi.org/10.37417/rivitsproc/3204

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*, 441-459. https://doi.org/10.1007/s11023-019-09502-w

Páez, A. (forthcoming). Transparencia, explicabilidad y confianza en los sistemas de aprendizaje automático. In J. D. Gutiérrez & R. Manrique (Eds.). *IA en acción: Desarrollo, impacto e implicaciones éticas en la sociedad*. Uniandes.

Santosuosso, A., & Sartor, G. (2022). La giustizia predittiva: una visione realistica. *Giurisprudenza italiana* (July 2022), 1759-1782.

Santosuosso, A., & Sartor, G. (2024). *Decidere con l'IA: Intelligenze artificiali e naturale nel diritto*. Il Mulino.

Simón Castellano, P. (2023). Taxonomía de las garantías jurídicas en el empleo de los sistemas de inteligencia artificial. *Revista de Derecho Político*, 117, May-August, 153-196.

Simón Castellano, P. (2023a). Allende una teoría general de las garantías jurídicas para una inteligencia artificial confiable. In F. Balaguer Callejón & L. Cotino Hueso, *Derecho público de la inteligencia artificial* (pp. 111-148). Fundación Manuel Giménez Abad.

Smuha, N. (2025). *Algorithmic Rule by Law: How Algorithmic Regulation in the Public Sector Erodes the Rule of Law*. Cambridge University Press.

Smuha, N. (forthcoming). Artificial Intelligence in the Judiciary: a Threat to the Rule of Law? In M. Zalnieriute & A. Limante (Eds.). *The Cambridge Handbook of AI and Technologies in Courts*. Cambridge University Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6264778

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, 11, 105-112. https://doi.org/10.1007/s10676-009-9187-9

Université de Montréal (2018). *The Montréal Declaration for a Responsible Development of Artificial Intelligence* https://montrealdeclaration-responsibleai.com/the-declaration/

Van Dijck, G. (2022). Predicting Recidivism Risk Meets AI Act. *European Journal of Criminal Policy and Research,* 28, 407-423 https://doi.org/10.1007/s10610-022-09516-8

Walters, J.; Dey, D.; Bhaumik, D., & Horsman, S. (2024). Complying with the EU AI Act. *Artificial Intelligence*. In S. Nowaczyk *et al.* (Eds.) *ECAI 2023 International Workshops. Kraków, Poland, September 30 – October 4, 2023 Proceedings, Part II* (pp. 65-75). Springer.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the Department of Law and the Faculty of Law at the Pontificia Universidad Católica del Perú for supporting my travel from Lima to Zagreb and for granting me a research leave. I am also grateful to Professors Sérgio Arenhart, Elisabetta Silvestri, Carlo Vittorio

Giabardo, Jordi Ferrer Beltrán, and Gabriella Valenzuela for their valuable insights and suggestions, as well as to the members of the PRODEJUS-PUCP Research Group for their continuous support.