The algorithm as judge: predictive justice, decision-making power and digital inequalities

by *Vera Kopsaj*[*], *Sara Sbaragli*[**]

Artificial intelligence is redefining judicial reasoning, discretion and legitimacy. Predictive justice tools influence risk assessment and sentencing, raising questions about fairness and accountability. This article takes a sociological perspective to examine the implications of algorithmic decision-making in justice. Drawing on legal theory, sociology of law and critical algorithm studies, we examine artificial discretion, opacity, and the reproduction of inequality through data. Based on models that consider algorithms as socio-technical actors, we examine how they redistribute power within digital institutions. We also address the augmented justice approach, which combines human judgement with algorithmic support. Our central question is how AI is transforming judicial decision-making in terms of discretion, legitimacy and justice.

Keywords: Justice systems; predictive justice; theory of law; sociology of law; algorithmic governance; AI and law.

L'algoritmo come giudice: giustizia predittiva, potere decisionale e disuguaglianze digitali

L'adozione dell'intelligenza artificiale nei sistemi giudiziari sta trasformando il ragionamento giuridico, la discrezionalità e la legittimità delle decisioni. Gli strumenti di giustizia predittiva, basati su algoritmi di apprendimento automatico, influenzano sempre più le valutazioni del rischio e le sentenze, sollevando interrogativi su equità e responsabilità. Questo articolo, da una prospettiva sociologica, analizza le implicazioni epistemologiche, etiche e politiche del *decision-making* algoritmico, esaminando discrezionalità artificiale, opacità del calcolo e riproduzione delle disuguaglianze. Si propone infine un quadro che interpreta gli algoritmi come attori socio-tecnici che codificano e ridistribuiscono il potere nelle istituzioni digitali. La domanda di ricerca è: quali trasformazioni introduce l'IA nel processo decisionale giudiziario in termini di discrezionalità, legittimità e giustizia sociale?

Parole chiave: sistemi giudiziari; giustizia predittiva; teoria del diritto; sociologia del diritto; governance algoritmica; IA e diritto.

DOI: 10.5281/zenodo.17297567

[*] UniCamillus - Saint Camillus International University of Health and Medical Sciences. vera.kopsaj@unicamillus.org.

[**] Università di Napoli "Federico II". sarasbaragli@gmail.com.

Introduction: From human judgment to algorithmic governance

AI is transforming the way legal decisions are made and justified. Systems such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in the US, used to predict the risk of reoffending, or Estonia's pilot programmes involving digital judges, signal the rise of algorithmic governance in the courts.

The COMPAS case shows how algorithmic tools encode uncertainty and bias. The Core Report of the COMPAS-R (2022) shows a predictive accuracy of approximately 0.68 AUC − only moderately better than the case − and significant disparities in performance between demographic groups (e.g. black women versus white women). Furthermore, research (Engel, Linhardt, Schubert, 2025) suggests that COMPAS structurally biases outcomes against defendants, promoting a precautionary rationale for detention, often reversing the presumption of innocence. Judges using COMPAS may prefer to detain people who are unlikely to commit a crime rather than risk releasing someone flagged as high risk, even though such predictions do not have strong statistical reliability. These critical issues are exacerbated by the proprietary and opaque nature of the tool, which limits public scrutiny and raises questions of democratic legitimacy.

This reflects a broader shift: decision-making is increasingly being entrusted to automated systems. Predictive justice is a socio-technical change that redefines how decisions are made, by whom, and under what conditions. As Bowker and Star (1999) remind us, classification systems are not neutral, but shape social life, define institutional boundaries, and assign visibility or invisibility to individuals and categories.

From the perspective of legal philosophy, this raises critical questions about the nature of judgement, interpretation and moral reasoning. Tuzet (2020) provocatively asks whether the algorithm becomes a "shepherd" of the judge. If discretion is historically understood as a space for interpretation and moral reasoning, algorithmic systems compress this space within probabilistic models, pre-structuring choices and expectations.

The delegation of evaluative authority to AI systems poses epistemic challenges: on what basis should we trust the results of systems we cannot fully understand? Legal actors often lack the technical expertise to assess the validity of machine-generated risk scores, which further complicates the dynamics of responsibility. As Mezza (2018) points out, the power of computation tends to create a cognitive shortcut, a perceived objectivity that obscures the incorporation of algorithms into social values, data policy and institutional priorities.

Moreover, algorithmic tools are being introduced in a context characterised by pressures for efficiency, cost reduction, and standardisation. These pressures make AI attractive to politicians and judicial administrators, especially in overburdened systems. However, as Sartori (2025) warns, we must remain vigilant against the enchantment of technology, adopting instead a critical attitude rooted in sociological disenchantment.

This evolution requires a profound rethinking of the very concept of justice. Is justice merely the accurate classification of cases, or is it an inherently interpretative, situated and human activity? The following sections explore these dilemmas, drawing on interdisciplinary studies to analyse the sociological consequences of the algorithm as judge.

1. Predictive justice and the delegation of responsibility

Predictive justice is based on machine learning trained on large datasets of past legal cases. The aim is to generate probability scores: the likelihood of reoffending, the level of risk or the probability of success in court. Although these tools claim to improve efficiency and objectivity, their epistemic basis is fragile and they often struggle to strike a balance between predictive accuracy and fairness, causing disparate impacts among protected groups (Barocas, Selbst, 2016). As Rundo and Di Stallo (2019) note, the algorithmic model is only as good as the data from which it learns.

Donati (2020) frames this shift as an evolution towards second-generation justice, in which human judgment is progressively replaced by calculated suggestions. Cominelli (2025) introduces the notion of “artificial discretion”: a hybrid logic in which the human role is maintained but strongly conditioned by the outputs of opaque systems. Judges thus find themselves in a paradoxical position: they have to answer for decisions partially made by others.

Further evidence from COMPAS underlines these concerns. According to the COMPAS-R Core Report (2022), risk scores are derived from variables such as previous arrests, age and socio-economic background, factors that may indirectly encode systemic bias. The reported accuracy (AUC around 0.68) highlights not only the limited predictive ability, but also the inherent uncertainty of these scores. Engel, Linhardt and Schubert

(2025) argue that this uncertainty is rarely communicated transparently to judges, which may lead them to overestimate the reliability of algorithm results.

This change makes accountability less clear and weakens legal reliability. Judges often rely on algorithmic risk scores because they have become institutionalised.

Moreover, the probabilistic nature of predictive instruments introduces a structural ambiguity. A 70% risk is not a fact, but a deduction, uncertain and subject to interpretation. Nevertheless, these tools often frame such results as authoritative or objective, fostering an epistemic conflict between probability and certainty in judicial decision-making (Rundo, Di Stallo, 2019; Donati, 2020).

This confusion between probability and certainty is amplified by the presentation of risk scores without sufficient explanation of their margin of error or bias. The lack of transparency, combined with institutional pressures to adopt such instruments, further limits the ability of judges to exercise informed discretion.

This ambiguity intersects with another key issue, that is the reduction of individuals to data profiles. The complexity of human behaviour is flattened into measurable variables, which are then fed into systems that produce predictions based on historical correlations. As Bevilacqua (2025) points out, this process of abstraction risks dehumanising justice, making it less sensitive to the singularities of each case.

In practice, the use of predictive tools may shift the logic of judgment from interpretation to classification, from deliberation to ranking. This reorientation may inadvertently promote a logic of punishment less concerned with individual culpability and more with actuarial risk management. This shift reflects broader transformations in contemporary governance, where the language of safety, efficiency and risk replaces that of rights, justice and equity.

Ultimately, predictive justice involves not only a technical delegation, but a normative transformation. It reconfigures what counts as relevant information, who is authorised to interpret it, and how decisions are legitimised.

Below, we explore the dynamics of predictive justice, examining how algorithmic systems operate as socio-technical actors and how their institutionalisation reshapes the distribution of power, trust and responsibility within the justice system.

2. A sociological perspective on algorithmic judgment

From a sociological perspective, algorithms should not be seen as neutral tools, but as actors embedded within socio-technical assemblages. As Artieri (2020) suggests, algorithms are cultural agents that classify, normalise and simplify social reality. They not only reflect social structures; they also produce them. The use of AI in the justice system is not an isolated technical update, but part of a broader institutional change in which power, trust and legitimacy are being reconfigured through technology.

Bevilacqua (2025) introduces the idea of digital normativity: a shift in the production of legal meaning from textual interpretation to data-driven modeling. This change involves the gradual replacement of language, precedents and hermeneutics with statistical correlation and pattern recognition. In this context, the authority of the algorithm emerges not from argumentative rigour or democratic deliberation, but from its metric performance, scalability and institutional approval.

The sociological relevance of algorithms lies precisely in their double nature: both technological artifacts and institutional scripts. They enact regimes of visibility and invisibility by determining what data matter, which characteristics are prioritised and which categories become actionable, what Fourcade and Healy (2024) describe as a logic of "ordinal society", in which people are classified, evaluated and governed accordingly. As Finco (2024) argues, this new communicative rationality shifts the focus from shared meanings to automated decision flows. The court, once a space of symbolic mediation, risks becoming a platform for procedural optimisation.

Joyce and Cruz (2024) emphasise the concept of data justice, highlighting the moral and political implications of decisions based on potentially partial or incomplete data sets. When trained on biased data, predictive tools risk perpetuating injustice, which Aanestad *et al.* (2021) describe as a form of digital inequality, where fairness and equal treatment depend on the quality and representativeness of the available data. There is also the challenge of opacity: algorithmic decisions often lack the transparency and accountability found in traditional processes, with proprietary systems limiting public scrutiny. Even experts can find it difficult to understand or contest the way risk scores are generated. In summary, the sociological analysis of algorithmic judgement reveals its performative nature, as algorithms reshape institutional justice logic and intersect with systemic inequalities to create new forms of exclusion and reinforce existing power structures.

3. Algorithmic bias and reproduction of inequalities

Algorithmic systems used in judicial contexts are only as neutral as the data on which they are trained. This fundamental limitation is well documented by empirical studies that show how predictive tools often perpetuate structural inequalities rather than eliminate them. Eubanks (2012), in her groundbreaking work on automated welfare systems, demonstrates how data-driven technologies disproportionately penalise marginalised populations, particularly low-income and racially diverse communities. Similar dynamics are at work in the judicial system.

Predictive algorithms rely heavily on historical data – arrests, sentencing patterns, demographic trends – that are themselves the product of systemic biases. As Sloane (2019) argues, inequality is not an unfortunate by-product of algorithmic reasoning, but rather a constitutive feature: "inequality is the name of the game". Algorithms trained on models of excessive surveillance or harsher penalties for specific groups inherit and reproduce those same discriminatory patterns. Their results may appear objective, but they are built on distorted foundations.

This becomes more problematic when feedback loops occur. For example, a risk assessment tool that flags individuals in certain zip codes as high risk may lead to increased surveillance and arrest rates in those areas. These new arrests feed back into the dataset, reinforcing the initial bias and creating a self-fulfilling prophecy. As Desrosières (1998) points out, quantification shapes the realities it claims to measure, incorporating institutional assumptions into classifications.

Lazar and Stone (2024) develop a theory of predictive justice, arguing that unequal predictive performance between structurally disadvantaged and advantaged groups constitutes a form of moral error, regardless of outcomes. Predictive models trained on historically unfair data risk representing certain populations epistemically incorrectly, thereby reinforcing structural inequality. This concern echoes what Galli and Sartor (2023) describe as the "digitalisation of deviance", whereby algorithmic classifications redefine the boundaries of suspicion and control.

Furthermore, the impact is not evenly distributed. Already marginalised groups are the most exposed to algorithmic decisions. Queudot and Meurs

(2018) warn against “differential visibility”: those who are most visible in data sets are the most controlled and punished, while privileged groups often benefit from algorithmic invisibility, as their data is less collected or used.

This reproduction of inequality is not simply technical, but reflects socio-political decisions. As Joyce and Cruz (2024) argue, data justice requires addressing the power dynamics behind data collection and use. A critical sociology of predictive justice must go beyond accuracy or fairness to examine the broader social impacts of delegating decisions to systems that amplify inequality.

This also raises urgent normative questions: Can a system be fair if it consistently produces unfair results? What forms of control or resistance are available to those affected? These questions point to the need to rethink the design, governance and accountability structures of AI in the legal sphere.

The next section explores possible frameworks for resisting algorithmic domination and claiming a human-centered vision of justice.

4. Disenchantment, trust and human supervision

The growing role of artificial intelligence in judicial systems presents a paradox: while intended to increase efficiency and fairness, algorithmic tools often generate opacity, alienation and distrust. This calls for renewed attention to human oversight and institutional accountability.

Sartori (2025) calls for “sociological disenchantment”, a methodological position that resists technological utopianism and emphasises the constructed nature of AI systems. Disenchantment is not a rejection of technology, but an invitation to question its assumptions, limitations and effects. When algorithms mediate decisions relating to liberty or punishment, they must be subject to ethical scrutiny just like human actors.

An urgent challenge is the erosion of transparency. Algorithmic decisions often lack intelligibility, making it difficult for defendants, lawyers, or even judges to understand how a score or classification was generated. This opacity threatens procedural justice and undermines public trust. As Barberis (2022) reminds us, justice must not only be done, but also seen to be done: decision-making must remain interpretable and contestable.

Barberis proposes a middle way: imagining algorithms as *auxiliaries* rather than *substitutes* for judicial reasoning. According to this model, AI provides analytical support without undermining the centrality of human discretion. To ensure this balance, institutions must implement robust oversight mechanisms, including algorithmic auditing, transparency standards and participatory design processes. These mechanisms can help restore a sense of agency to legal practitioners and affected individuals.

Education is also crucial. Mangone, Martini and Volterrani (2025) emphasise the importance of new training programmes that equip lawyers, policymakers and citizens with the critical skills needed to interact with algorithmic systems. This means not only technical literacy, but also ethical reasoning, legal imagination and sociological insight.

Trust in justice cannot be programmed, but must be earned through responsability and human engagement. This implies redefining professional roles in hybrid environments where legal reasoning and algorithmic results coexist, but are not confused with one another.

Ultimately, disenchantment is not cynicism, but responsibility. It affirms the need for of a human-centered approach to justice that recognises the potential of AI, while insisting on the irreplaceable value of human judgment. The next section explores how such an approach might be realised in the context of augmented justice.

5. Toward augmented justice?

The risks described above highlight the need to rethink how artificial intelligence is integrated into judicial institutions. Rather than adopting a binary view – man versus machine, discretion versus automation – we should adopt the existing framework of augmented justice, which emphasises the complementarity between humans and AI in legal decision-making.

Picierno (2025) calls for a "hybrid legal culture" in which judges are supported, not replaced, by computational systems. This vision requires infrastructures that enable collaboration under conditions of transparency, contestability and ethical control. Augmented justice is based on interpretability, contextualisation and a commitment to prioritise justice over efficiency.

Galli and Sartor (2023) outline models of judicial empowerment: diagnostic, advisory, and decision support. The most effective uses of AI lie in the first two: providing insights into case patterns, relevant precedents, or inconsistencies, while the final interpretation remains the responsibility of the judge. This preserves discretion, empathy, and normative judgement.

Furthermore, augmented justice requires new forms of participatory governance. Citizens, legal practitioners, civil society organizations and technical experts must be involved in the design, evaluation and regulation of AI systems used in the justice sector. This is in line with democratic ideals and ensures that algorithmic systems remain accountable to the public they serve. As Dyson (1997) argued decades ago, technology must be embedded in social justice structures: it is not inherently liberating or oppressive, but is shaped by the way we choose to govern it.

Augmented justice is not a technical solution to a technical problem. It is a political and normative project that recognises the complexity of human judgement and the social embeddedness of legal institutions. Queudot and Meurs (2018) warn against techno-solutionism in legal innovation, reminding us that algorithms cannot resolve the deeply rooted inequalities and ambiguities that characterise legal practice. Rather than eliminating discretion, augmented justice reaffirms it as a locus of human responsibility.

In this sense, the future of justice is not algorithmic, but augmented: informed by data, assisted by computation, but ultimately implemented by human beings. The ethical imperative is to design legal technologies that enhance, rather than diminish, the quality of judgement and to foster institutional ecosystems capable of critically engaging with the tools they use.

Conclusion

The spread of artificial intelligence in judicial systems marks an important institutional change. Predictive justice redefines discretion and legitimacy and can exacerbate inequalities when personal dignity is ignored. Algorithms act as socio-technical agents: they classify, redistribute authority and can amplify existing biases.

We have demonstrated how opacity, probabilistic logic and data-driven reasoning affect fairness and legitimacy. Without critical scrutiny, such systems risk undermining procedural justice and public trust.

We build on the augmented justice model, a framework that envisions hybrid systems where AI supports, not replaces, human judgment. It prioritises transparency, ethical design, education and participatory governance to ensure discretion and accountability.

Future research may explore tools such as COMPAS and how to access fairness. Surveys of judicial actors could further clarify how these systems are understood and used in practice. Empirical investigation will be key to refining augmented justice and promoting socially responsible innovation.

References

Aanestad M., Kankanhalli A., Maruping L., Pang M.S., Ram S. (2021). Digital technologies and social justice. *MIS quarterly*, *17*(3): 515-536.

Artieri G.B. (2020). Fare Sociologia attraverso l'algoritmo: potere, cultura e agency. *Sociologia italiana*, (15).

Ayuda F.G. (2024). Algorithms, sociology of law and justice. *Journal of Digital Technologies and Law*, *2*(1): 34-45.

Barberis M. (2022). Giustizia predittiva: ausiliare e sostitutiva. Un approccio evolutivo. *Milan Law Review*, *3*(2):1-18.

Barocas S., Selbst A.D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bevilacqua S.A. (2025). Gli strumenti di intelligenza artificiale nel sistema giudiziario: verso una nuova normatività digitale. *Sociologia del diritto*, 52(1): 257-274.

Bowker G.C., Star S.L. (1999). *Sorting things out: Classification and its consequences*. The MIT Press.

Cominelli L. (2025). Discrezionalità artificiale e giudizio algoritmico: Artificial discretion and algorithmic judgment. *Sociologia del diritto*, 52(1): 242-255.

Cordano M. (2025). Intelligenza artificiale e giustizia: una valutazione critica del caso Loomis vs. Wisconsin.

Desrosières A. (1998). *The politics of large numbers: A history of statistical reasoning*. Cambridge: Harvard University Press.

Donati F. (2020). Intelligenza artificiale e giustizia. *RIVISTA AIC* (415-436).

Dyson F. (1997). Technology and social justice. *Carnigie Concil on Ethics and International Affairs* (7-25).

Engel C., Linhardt L., Schubert M. (2025). Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. *Artif Intell Law*, 33: 383-404. https://doi.org/10.1007/s10506-024-09389-8

Eubanks V. (2012). *Digital dead end: Fighting for social justice in the information age*. Cambridge: MIT Press.

Finco M. (2024). Dall'Intelligenza alla Comunicazione artificiale? Sociologia e possibilità teoriche. *indiscipline-rivista di scienze sociali*, *4*(2): 132-140.

Fourcade M., Healy K. (2024). *The ordinal society*. Cambridge: Harvard University Press-T.

Galli F., Sartor G. (2023). Ai approaches to predictive justice: A critical assessment. *Humanities and Rights Global Network Journal*, *5*(2).

Joyce K., Cruz, T.M. (2024). A sociology of artificial intelligence: Inequalities, power, and data justice. *Socius*, *10*, 23780231241275393.

Lazar S., Stone J. (2024). On the site of predictive justice. *Noûs*, *58*(3): 730-754.

Mangone E., Martini E., Volterrani A. (2025). *Società frammentata e traiettorie educative: Disuguaglianze, giustizia sociale e intelligenza artificiale*. Milano: Mimesis.

Mezza M. (2018). *Algoritmi di libertà: la potenza del calcolo tra dominio e conflitto*. Roma: Donzelli Editore.

Northpointe Inc. (2022). *COMPAS-R Core: The revised version of the standard COMPAS Core (Technical manual).* Research Department, Northpointe, Inc. https://github.com/franzgualdi/PA-Algorithmic-formalization/blob/main/COM PAS-R%20Core%20Research%20Report%203%2028%2022%20(1).pdf

Picierno B. (2025). Giustizia predittiva ed intelligenza artificiale: nuovi scenari per il giurista del futuro. *Futuri*, *23*: 361-372.

Queudot M., Meurs M.J. (2018). Artificial intelligence and predictive justice: Limitations and perspectives. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (889-897). Cham: Springer International Publishing.

Rundo F., Di Stallo A.L. (2019). Giustizia predittiva: algoritmi e deep-learning. *Sicurezza e Giustizia*: 31-34.

Sartori L. (2025). Conoscere e governare la tecnologia per essere più disincantati di fronte all'intelligenza artificiale. *Rassegna Italiana di Sociologia*, *66*(2): 397-405.

Sloane M. (2019). Inequality is the name of the game: thoughts on the emerging field of technology, ethics and social justice. *Weizenbaum conference*. DEU.

Tuzet G. (2020). L'algoritmo come pastore del giudice? Diritto, tecnologie, prova scientifica. *Media Laws*, (1): 45-55.