laws

MDPI

Article

AI Accountability in Judicial Proceedings: An Actor–Network Approach

Francesco Contini [1,*], Elena Alina Ontanu [2,*] and Marco Velicogna [1,*]

[1] Institute on Lega Informatics and Judicial Systems, National Research Council of Italy, 40125 Bologna, Italy
[2] Tilburg Law School, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands
* Correspondence: francesco.contini@cnr.it (F.C.); e.ontanu@tilburguniversity.edu (E.A.O.); marco.velicogna@cnr.it (M.V.)

Abstract: This paper analyzes the impact of AI systems in the judicial domain, adopting an actor–network theory (ANT) framework and focusing on accountability issues emerging when such technologies are introduced. Considering three different types of AI applications used by judges, this paper explores how introducing non-accountable artifacts into justice systems influences the actor–network configuration and the distribution of accountability between humans and technology. The analysis discusses the actor–network reconfiguration emerging when speech-to-text, legal analytics, and predictive justice technologies are introduced in pre-existing settings and maps out the changes in agency and accountability between judges and AI applications. The EU legal framework and the EU AI Act provide the juridical framework against which the findings are assessed to check the fit of new technological systems with justice system requirements. The findings show the paradox that non-accountable AI can be used without endangering fundamental judicial values when judges can control the system's outputs, evaluating its correspondence with the inputs. When this requirement is not met, the remedies provided by the EU AI Act fall short in costs or in organizational and technical complexity. The judge becomes the unique subject accountable for the use and outcome of a non-accountable system. This paper suggests that this occurs regardless of whether the technology is AI-based or not. The concrete risks emerging from these findings are that these technological innovations can lead to undue influence on judicial decision making and endanger the fair trial principle.

Keywords: AI systems for justice; accountability; actor–network theory; AI Act; speech-to-text applications; legal analytics systems; predictive systems

Citation: Contini, Francesco, Elena Alina Ontanu, and Marco Velicogna. 2024. AI Accountability in Judicial Proceedings: An Actor–Network Approach. *Laws* 13: 71. https://doi.org/10.3390/laws13060071

Academic Editor: Patricia Easteal

Received: 1 July 2024
Revised: 6 November 2024
Accepted: 15 November 2024
Published: 23 November 2024

1. Introduction

The rumbling development of artificial intelligence (AI) is affecting any professional area with new applications and the promise of deep changes. In the legal sector, AI systems already provide a variety of functions including predictive justice, anonymization, automatic translation, speech to text, legal analytics (Reiling 2020), and automatic drafting (Pierce and Goutos 2024), just to mention a few. Besides the numerous envisaged benefits, AI introduction also poses legal, organizational, and ethical challenges, including potentially disruptive changes to judicial decision making (Kyriakides et al. 2022, p. 121; Velicogna 2007; Zuckerman 2020, p. 427). It is unquestioned that IT systems, particularly those based on AI, suffer due to limited transparency, accountability, and lack of explainability, which are pre-requisites of key importance in justice proceedings (Goodman and Flaxman 2017; Pégny et al. 2019). When non-accountable and explainable technologies influence judicial procedures and decisions, they pose the risk of undue influence and hence impact the fair trial principle, particularly regarding judicial independence and the equal treatment of justice seekers (Angwin et al. 2016; Contini 2020; Galdon-Clavel et al. 2024). Relevant efforts have been made to build and increase AI accountability. Nevertheless, AI remains essentially non-accountable (Galli and Sartor 2023).

In the process of waiting for technological fixes for the problems of accountability and explainability, this paper elaborates on the impact of introducing systems which are neither accountable nor explicable in judicial proceedings. From this standpoint, it assesses the impact of non-accountable and inexplainable applications on the judges' activity. To carry this out, it focuses on the comprehensive accountability of the technology and human constellation by following the actor–network approach (Latour 2005; Monteiro 2000).

Human actors, judges in our case, rely on technological systems to deliver or support their working process. In this, there is not an accountability problem when humans and technology can be held accountable for their respective actions. This is the case of automated case management systems. With this kind of applications, it is possible to identify errors and understand if an error can be attributed to design issues, technical glitches or to human mistakes. What happens, then, when non-accountable systems are introduced in the actor–network? This paper shows how, inevitably, human actors maintain accountability for outcomes, even when not in control of the system's outputs or of the explanation about how an outcome has been reached. Is this acceptable in the judicial domain? To what extent can AI-type applications support judicial tasks without undermining key judicial values? What organizational measures can be taken to accommodate the use of such technologies?

This paper addresses this topic from an EU perspective on technology for justice. The EU provides a unique legal and international framework and a recently adopted AI legislation which applies to its Member States. Furthermore, in the EU, the protection of fundamental rights benefits from a strong legal and institutional base and practice under the provisions of Article 6 of the European Convention of Human Rights (ECHR) and the case law of the European Court of Human Rights, as well as Article 47 of the European Charter of Fundamental Rights of the European Union (the Charter) and the case law of the Court of Justice of the European Law.

To understand the context in which this analysis is further grounded, an overview of European legal developments is succinctly provided in Section 2. Section 3 explains the methodological choices made to answer the identified questions. For this, three categories of AI-based systems with different configurations of the interaction between technology and the human agency have been chosen to explore the impact technology can have on human actions and accountability. The actor–network theory (ANT) is introduced in Section 4. Then, Sections 5–7 each focus on the different categories of AI systems that can support justice proceedings (i.e., speech to text, legal analytics, and predictive justice) by considering the relationships between technology and humans (the judge and other subjects) and how technology agency can impact accountability. Section 8 analyzes whether and how humans can understand technology-based agency, assesses to what extent the new actor–network configuration affects judicial accountability, and assesses how the technological and organizational arrangements can be designed to keep the system aligned with fair trial requirements

2. An Overview of the General EU Reference Framework

For over a decade, the EU and other international institutions have actively encouraged and pursued digitalization at various levels of justice systems (e.g., the use of technology in court proceedings, dedicated online portals, and the digital handling of European procedures), seeking to improve different areas of the administration of justice. The European e-Justice Strategy and the 2019–2023 Action Plan set the use of AI and distributed ledger technology as priority areas in the justice field. In the 2019–2020 period, the EC promoted a study to map out their deployment (Spasojevic et al. 2020, p. 31). References to AI are also made in the latest 2024–2028 European e-Justice Strategy where AI use is envisaged among actions that can contribute to facilitating access to justice and enhance the efficiency and effectiveness of justice services in the future. Although for the time being, the concrete uptake of AI within courts remains on the low side, particularly for systems predicting decision, supporting legal research, or easing judicial drafting (CCJE 2023), interest in developments and integration of AI in the justice domain is high.

In accordance with Article 2 TEU, the use of technology, including AI, must be in line with common EU values: respect for fundamental rights as stipulated by the Charter and the ECHR, freedom, democracy, equality, and the rule of law. This gives rise to an accountability issue, that is, how to ensure that any deployment of AI-based systems is compliant with these fundamental human rights and core EU values.

More generally, after a first wave of enthusiasm, pitfalls of AI technologies and the risks connected to its use pushed EU institutions to approve, on 21 May 2024, the so called "AI Act", the first regulative framework to control and limit the use of AI systems. Regulation (EU) 2024/1689 laid down harmonized Member States rules on AI and established requirements and obligations for AI developers and deployers regarding specific uses of AI. The AI Act (Article 3(1)) defines an AI system as a "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." All AI technologies identified based on this very broad definition are then classified into four different levels of risk, ranging from minimal or no risk to unacceptable. Justice-related developments are mostly expected to be high-risk AI systems (see point 8 of the Annex III AI Act) unless they involve purely ancillary administrative activities that do not impact the actual administration of justice. The driving idea is regulating AI to prevent harmful outcomes. AI systems should be "overseen by people, rather than by automation" (European Parliament 2024).

3. Methodology

The methodology identifies AI applications with different risk levels in accordance with the AI Act. Even if all systems used in the justice domain are considered high risk (with the exception of those supporting ancillary administrative activities), different applications show different risk profiles. Speech-to-text technology used for judicial writing shows a risk lower than legal analytics, which propose to the judge relevant jurisprudence with the help of AI algorithms. A higher risk is associated with predictive systems that suggest to the judges the decisions they should take. Moving from this selection of the type of applications, the following step is to identify concrete examples of AI systems used by judiciaries. The search of these cases included the 2020 European Commission Study on the use of innovative technologies in the justice field (Spasojevic et al. 2020), and the more recent CCJE 2023. These works show that while the EU has developed the most comprehensive AI legal framework, its practical applications for justice are still quite limited. However, these works allowed the authors to identify two of the three relevant applications for this study. These include a variety of commercial speech-to-text applications to replace typing in many countries and the Smart Sentencing Project, a legal analytics application developed in Germany (Rostalski et al. 2021).

This approach, however, was not successful in the case of predictive justice. When searching outside the findings of CCJE and EU surveys, a relevant development was identified, namely the RisCanvi system used by prisons in Catalonia, Spain, to estimate the risk of inmates reoffending upon leaving prison (Bellio 2021). This system has similarities with the much better known COMPAS system used in the US. Hence, the analysis of sentencing support systems relies on the well-known US project—COMPAS—for which data and information are more extensive and accessible for research purposes, and on the Spanish RisCanvi system.

Distinguishing between these three groups of technologies allows us to observe different configurations of technology and humans, how technological agency affects human actions in the actor–network, and concrete mechanisms through which organizational arrangements can mitigate the risks associated with AI use in practice. In this endeavor, the actor–network theory (ANT) provides the main theoretical framework (Bijker and Law 1992; Czarniawska and Joerges 1998; Czarniawska 2004; Latour 2005, pp. 54–55). Furthermore, the analysis benefits from a multidisciplinary approach in the sense that it

will follow the legal perspective as well as perspectives of social sciences, sociological, and managerial studies wherever necessary, situating this paper at the crossroads of law and technology and law and society scholarship.

4. The Actor–Network Theory and the Question of Accountability

The actor–network theory (ANT) is relied on to explore how the introduction of non-accountable technological applications impact the socio-technical systems in which judges and AI systems interact to carry out a task. From the ANT perspective, law and technology are omnibuses for two different kinds of agency through which they influence human behavior. Specifically, (1) law exerts influence through prescriptive norms that delineate permissible and impermissible actions, while (2) technology affects behavior by offering structured workflows that guide actions and decisions or by automating tasks once performed by human agents (Bijker and Law 1992). Hence, the introduction of digital innovation in such a highly regulated field as the judiciary entails a transfer of agency from formal regulations (laws and other formal norms), pre-existing routines, and other organizational constraints to technological systems (Contini and Lanzara 2014, pp. 4–6).

ANT considers technologies as non-human actants emphasizing the capacity of artifacts—even simple ones—to act and influence humans. The action occurs in actor–network frameworks in which human agents and non-human actants (law, technology, and other organizational constraints) interact in various ways (Czarniawska and Joerges 1998; Latour 2005). In the judicial domain, a case management system executes actions previously performed by humans, such as summoning case parties and suggesting which operations can be undertaken in proceedings or recording the duration of events (Velicogna 2023). Consequently, the introduction of new technological artifacts changes the pre-existing actor–network state. First, some activities previously performed by humans are delegated to machines. Second, machines (algorithms or software) tell humans what to do, when, and how to carry out the activities concerned (Lanzara 2009, pp. 9–12). AI systems suggest how speech can be transformed into written words, which previous judgements are relevant for a present case, or which decision should be taken in a given proceeding. In a nutshell, the knowledge embedded in working practices and commands of formal regulations are "inscribed" into software codes. Through this process, the execution and the enforcement of judicial proceedings are delegated to software codes that become self-enforcing artifacts. Indeed, working against the constraints, affordances, and suggestions provided by technology can be impossible or demanding in terms of costs, time, or personal engagement (Garapon and Lassègue 2018; Lessig 2007). The consequences of such inscriptions are automation and guidance: the machine performs actions previously performed by humans and guides humans in the executions of judicial tasks. The transfer of agency associated with technological innovation has implications that are at the core of this paper: when agency is transferred to non-accountable AI, what happens to the broader accountability requirements of fair procedures and decisions?

The concept of accountability has been extensively debated for both judiciaries (Atchison et al. 1999; ENCJ 2018; Seibert-Fohr 2012) and technologies (Association for Computing Machinery and US Public Policy Council 2017; Chiao 2019; Johnson 2004; Nissenbaum 1996, 2007; Shah 2018). To integrate the two debates, we rely on Herbert Simon and colleagues' classical definition of accountability: "the combination of methods, procedures and forces determining which values are to be reflected in administrative decisions" (Simon et al. 1961, p. 513). It involves an obligation to inform and justify one's conduct to a third party (Novelli et al. 2024, p. 1872), which is a pre-requisite to check to what extent proper values are built into procedures and decisions.

In the case of judicial proceedings, technologies are one of the methods (or forces) that—together with law-based procedures and other mechanisms—contribute to defining the values reflected in processes and decisions. The concepts of inscription and delegation of law-based actions into technological apparatus show how technology—bringing into the system its specific agency—impacts, in several ways, the values reflected in procedures

and decisions. More precisely, law and technology should work together with the objective of achieving fair procedures and decisions according to the criteria established by national constitutions and international standards starting with Art. 6(1) ECHR. AI applications can improve access, efficiency, and, maybe, decisional consistency, but it can also challenge fairness or introduce biases and undermine equal treatment due to reliance on the group behavior approach rather than specific individual-focused behavior (McGregor et al. 2019). Accountability can thus be considered as a two-channel mechanism. It includes the arrangements which instill proper and relevant values, and those that make it possible to assess whether the organization builds those values and interests into its own procedures and decisions.

From this angle, the focus of this study is not on AI accountability per se, but on its impact on judicial accountability and its underpinning judicial values, and hence, on the actor–network configuration composed of technology actants and human agents that perform judicial functions. To explore this issue, the following sections assess the impacts of three types of technological systems with different risk profiles on the actor–network framework designed to handle judicial proceedings.

5. Speech-to-Text Applications: Individual Accountability

Embedded into a growing range of technological devices, from word processors to mobile phones, speech to text is the quintessential example of ubiquitous AI functionality[1]. These new systems speed up the process of writing procedural documents and complements the usage of the computer keyboard. While for the keyboard, the link between the pressure on a key and a sign on the screen is automatic and straightforward, the functioning of a speech-to-text algorithm remains inscrutable (Mittelstadt et al. 2016). In the judicial field, speech to text is used to delegate to machines the drafting of potentially any procedural document from its oral enunciation: court hearings, sentences, briefs, and petitions. However, the outcome of a speech recognition system is just a text without any procedural value. It becomes a procedural act when processed according to legal requirements, such as signatures and/or stamps (handwritten or electronic). The user, transforming the text into a legal deed through a process of validation which includes adding a signature and a stamp, becomes accountable for the correctness of the action performed by the technological system. Drafting with speech-to-text is not an issue even if the functioning of the technology is totally inscrutable, since the user can easily check whether the output generated by the machine correctly corresponds to the input. Looking at the actor–network configuration, the agency related to the "typing" of the document is transferred to the technology, while the users remain accountable for the output. In other words, the transfer of agency of the action to the algorithm is decoupled from the transfer of accountability, which remains with the user. In this specific case, this decoupling does not undermine the overall accountability of the actor–network configuration. The user of the speech-to-text technology can easily check if the output is correct even if the functioning of the application is totally non-accountable. A well-designed use of such systems can even increase the overall accountability of the procedure and ensure a fair trial. In many Italian courts, judges use speech-to-text applications to dictate minutes during the hearing. A second screen allows lawyers and parties to check, in real time, the output of the application, spot errors, and ask for amendments. The crosscheck made by the judge, the lawyers, and parties zeros the risk of having the content of the record disputed: AI—even if not accountable—is used in an organizational setting which increases transparency and efficiency and protects fairness and the due process of law (Contini 2020).

Machine accountability is not an issue when technology is used in this way. Even an "inscrutable algorithm" can improve the delivery of justice when coupled with effective forms of human supervision and accountability. At the same time, the complete lack of

[1] See, for instance, the description of Google speech-to-text applications available in Google, "Speech-to-text" https://cloud.google.com/speech-to-text/ (accessed on 24 May 2024).

technological accountability does not impede the assessment of the values which must be built into the procedures. As the case shows, the use of AI, even if non-accountable, is not a problem per se for this type of developments.

6. Legal Analytics: Towards Collective Accountability

Legal analytics and, more generally, computational legal studies, have existed for a long time, certainly before AI. Hence, experiences of legal analytics include those based on "traditional" digital technologies with systems that extract knowledge (i.e., structured data or metadata) from the text of sentences and other legal documents that are then elaborated to ease the search of relevant information or to conduct an analysis based on the different socio-legal frameworks (e.g., statistics, legal analysis, etc.). The contemporary AI turn made it possible to delegate to the machine a growing chunk of tasks related to knowledge extraction, management, and the analysis of information. This delegation can reduce the costs and time needed to carry out the task and makes it possible to analyze growing bodies of legal texts within short timeframes. The comparison of two different systems, one based on traditional digital technologies and substantive human inputs, and the other on machine learning and other AI technologies, helps to clarify the different challenges and further discuss the issue of accountability.

Databases of judicial decisions are quintessential elements of any judicial system since they are constitutive components of the corpus juris of a jurisdiction. Traditionally collected in judgments' books and indexed through simple indexes, they are one of the first artifacts to be digitalized in the judicial landscape. Since the 1970s, Supreme Courts' judgements have been first indexed in electronic databases using the pre-existing criteria, sometimes supplemented by other entries like keywords. Lately, the same judgments have been collected in an electronic format. Indexes allowed for data retrieval through simple queries searching or filtering the indexes and—when available—in the full text of the decisions. The retrieval procedures extract the sentences by exclusively following the simple research criteria mentioned above, and the results are ordered following what was requested by the query. The reply of the machine is mechanical.

For instance, Sentenze Web, the Italian Court of Cassation databases, allows for judgments to be filtered based on the matter (civil or criminal), the type of decision (sentence, court order, or decree), the chambers of the courts, and the year. All of this follows a pre-established XML schema based on international standards.[2] The coding of the XML, which makes the search possible, is made by the Electronic Documentation Center of the Court of Cassation. Then, the search can be carried out through keywords, the decision number, law, or regulative reference. The results are provided in chronological order, hearing date, or relevance. The work of classification is mechanical, with little if no discretion, and is carried out within or under the supervision of the Court of Cassation. Software developers and those engaged in the coding of XML files have inscribed into the system the data and search criteria. They have transferred into the system the agency required to search relevant jurisprudence. Sentenze Web (and its developers) are accountable for the functioning of the system and its ability to provide the complete set of Court of Cassation jurisprudence and straightforward search tools. Users are accountable for the queries they make and for the use they make of the information extracted. The transfer of agency to a technological system is coupled with a transfer of accountability. The Court of Cassation and its Data Center are accountable for the features of the system (lack of bias, completeness, lack of manipulation, or nudging), and the users are accountable for their use of the information (decisions) retrieved. This approach, which should be understood as automatic and mechanical, is not followed when AI technologies are introduced to extract decisions or to analyze a set of judgments. As the approach becomes probabilistic and based on "soft logics" (Kirsch et al. 2020), keeping the judiciary in control becomes more challenging.

[2] See http://www.akomantoso.org (accessed on 24 May 2024).

Sentenze Web is not a system that would be categorized as an AI system according to the AI Act given it is not meant to be an adaptative system that is autonomous from human intervention, nor does it generate predictions or suggestions for sentencing.

Going from traditional digital technologies to AI, a new set of opportunities and challenges emerges. Text analytics and text mining lead to an automation of the knowledge representation process by relying on "a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation" (Hu and Liu 2012). Legal analytics are based on the same process, thus deriving substantively meaningful insights from legal data, including legal textual data (e.g., legal provision and texts of judicial decisions). In comparison to Sentenze Web, data extraction is carried out by machines with initial support from staff dedicated to the definition of the data extraction algorithms and to the "training" of the system. The algorithms are trained to "identify patterns in data, summarize the patterns in a model", and hence, to analyze the data (Ashley 2017, p. 23). In comparison to a classical electronic collection of judgements, such as Sentenze Web, AI-empowered legal analytics not only provides dedicated queries to access the relevant corpus juris, but also generates outcomes in terms of analysis. With such empowered functions, the legal analytics programs that seek to provide support for judges in the decision-making process promise to increase efficiency and the fairness of the law. The aim is to improve the systematicity, transparency, and uniformity of judicial decisions given that it was found that sentencing and interpretation of the law can vary significantly between courts in the same region and/or various regions in the same country. By relying on such a program, developers hope to improve the work of judges, the trust of users in the judiciary, and the communication between courts and society.

Smart Sentencing is a system that analyzes criminal sentencing practices to inform judges and legal practitioners on the development of case law across time and jurisdictions of German courts.[3] Furthermore, since the system could be used to provide an overview on the interpretation of the law based on available court case laws, and its outputs can be used as references for sentencing new cases, it affects judicial decision making. Hence, it should be considered high risk. Furthermore, the development of the system encountered obstacles ranging from the incomplete collection of judgements to the inconsistent structure of the judgements and the lack of a classification system.

In the classification of the relevant corpus juris, the Smart Sentencing project approach is different from the one used by Sentenze Web. Its approach is comparable to that of other supervised machine learning systems (Dhungel and Beute 2024). A large amount of data had to be analyzed and classified by the researchers to enable the system to subsequently autonomously classify new judgements. In the initial process of data coding, special attention was given to each element that could have an impact on the judgment, even information that was not explicitly considered in the individual sentencing such as the socio-economic status of the accused (e.g., income, number of children, and criminal record). According to the developers, the standardized recording of reasons in Smart Sentencing contributes to the transparency of existing differences in judicial practice and promotes coherence in law application and judicial decision making (Dhungel and Beute 2024; Rostalski et al. 2021).[4] In the initial stage, the machine learning was not supervised by humans. This stage was followed by a guided learning process in which the researchers were tagging and labeling the correct results (Rostalski et al. 2021). Thus, a human "trainer"

[3] Rostalski et al. (2021). Frauke Rostalski, Smart Sentencing, Conference organised by the German Presidency of the European Council, *Digitalisation of Justice—Interconnection and Innovation*, 8 December 2020; Legal Tech Lab Cologne, Das Legal Tech Lab Cologne blickt zurück auf das erste Jahr seines Bestehens, 19 March 2020 (available at https://jura.uni-koeln.de/en/fakultaet/zentrale-einrichtungen/studien-und-karriereberatungszentrum/newsletter/april-2020/news-april-2020/rueckblick-des-legal-tech-lab-cologne, accessed on 24 May 2024).

[4] See also Legal Tech Lab Cologne, Das Legal Tech Lab Cologne blickt zurück auf das erste Jahr seines Bestehens, 19 March 2020 (available at https://jura.uni-koeln.de/en/fakultaet/zentrale-einrichtungen/studien-und-karriereberatungszentrum/newsletter/april-2020/news-april-2020/rueckblick-des-legal-tech-lab-cologne, accessed on 24 May 2024).

had to check and adjust the outputs of the AI-enabled system, following a supervised machine learning approach. The more complex the answers the program needs to give, the more training is required (e.g., in judgments involving multiple defendants and crimes, the AI has to be taught which type of information belongs to which crime, connect them correctly, and make relations between the various sets of data). Once the training is completed, the recognition becomes sufficiently precise such that the labeling of correct results is no longer needed, and the classification stages can be automatically executed by the algorithm that extracts relevant data from completely unknown legal text (i.e., new decision). At this stage, the human trainers have ended their work, and the machine is supposed to extract and classify data autonomously. Thus, the function of extracting relevant data from the body of sentences is carried out through two different actor–network arrangements in which the roles of humans (researchers and judges) and technology (algorithms and neural networks) are combined in various ways.

The program can classify decisions accordingly in typical case constellations and deviations from the general trend of sentencing.[5] Smart Sentencing can recognize data and process high volumes of judgments in a short timeframe. This means that the database can be continuously updated. The update is placed in the hands of the judges who will have to upload the texts of their decisions. This is intended to take only a few minutes per judgment and involves three steps: uploading the decision, controlling the classification automatically proposed by the machine, and, if necessary, making the corrections needed. The system is set to suggest the classification of each case based on a legal analytics process. In this way, the researchers, developers, court staff, judges, or trained staff do not need to carry out the categorization for each individual case to be fed into the legal analytics, but the judge retains control of the classification and, if necessary, can correct it. This can eliminate mistaken outcomes that can introduce biases or the inability of dealing with cases that are not part of the general categories the system is familiar with.

In Smart Sentencing, the judges can supervise the input in the system and the categorization. From the perspective of the actor–network architecture, the judge retains the final check of data quality—including the input of the decisions and the classification analysis carried out by the AI system—as well as the accountability for these two stages. In this regard, the accountability which could be attributed to the machine disappears because the agency attributed to it is limited to "a suggestion" of the classification subject that has to obtain the approval of the human agent. If properly executed, this judicial oversight can guarantee a quality check of the classifications automatically suggested by the machine.[6] However, it assumes judges will execute an additional task on the top of their already busy agenda. If they are not sufficiently motivated to carry out this task, there is a high risk that the sentences will be uploaded without adequate judicial control or not uploaded at all. If this happens, the sentences will be de facto classified without human (judicial) supervision. This poses a threat to the accuracy of the classification that is initially made by professional judges. Furthermore, the inner functioning of the machine remains obscure, and the individual judge is limited in checking whether the algorithms extracting the relevant sentences work as they should.

Thus, three additional issues need to be considered: (1) the consistency of the case classification by all judges and/or staff involved in the selection and adjustment of the classification proposed by the algorithm; (2) the identification of the introduction in the systems of irrelevant or undesirable elements that affect the accuracy of the suggestions made, although the classification of the case is not incorrect; and (3) the risk of the "*effet moutonnier*" (Garapon and Lassègue 2021) in which the judges and/or staff involved approve the suggestions of the algorithm without performing a real check of the suggestions. As with Sentenze Web, the use of Smart Sentencing does not aim to carry out the automation of sentencing but seeks to achieve transparency and predictability of the penalty. Judges,

[5] Ibidem.

[6] On the importance to maintain human oversight as well as the accuracy and robustness of the system, see Articles 14, 15, and 26(2) and Recital 73 in EU AI Act.

lawyers, and scholars can carry out research related to specific types of cases (e.g., issued by certain court(s), within a certain timeframe, and with a certain outcome) (Dhungel and Beute 2024). Together with this, the system provides the users with a map of courts' case law orientations regarding the queried legal matter. Besides the map and some information on the average of the case law outcome, the users can consult in parallel the texts of the decisions (Rostalski et al. 2021). Research conducted in Germany indicates that judges would be interested to use such database system if made available to them for supporting their assessment work in a case (Dhungel and Beute 2024).

As the system evolves due to being driven by the algorithm's development and individual judges' collective input, the outcomes of the queries evolve as well in suggesting the relevant decisions for specific situations and the outcomes of newer decisions. The suggestions given in relation to the main case law orientation may pose a threat to judicial values, especially judicial independence. This can be the case if the suggestions given by the system are used as such by a judge without a further analysis in relation to the individual facts and legal profiles of the case being assessed or if these suggestions are wrong or abusive compared to the factual situation the judge has to decide upon. Given these critical aspects, extensive access to information for judges was considered desirable by the Smart Sentencing project researchers. This allows the judges to reflect on societal values that are revealed in the practice of sentencing. Additionally, if properly used, the possibility to analyze the details of the cases identified by the queries allows the judges to maintain their independence and control over data and the way they rely on the acquired knowledge in the decision-making process. In coming to a decision in a new individual case, the system permits a differentiation between the individual characteristics of previous decisions[7] that are relevant for the analysis, leading to the formation of a judge's conviction and the specific facts the judge is called to assess in a new decision.[8] Being able to differentiate between individual characteristics of previous decisions is a point of attention that must be taken into consideration in the deployment of such software supporting judicial decision making and "fair sentencing"[9]; otherwise, issues of wrong categorizations of cases and/or unrelated or undesirable elements introduced in the suggested selection of relevant case law for decision making may be very difficult to spot and may be difficult for the human actor to properly address and correct. In such circumstances, the judge continues to maintain accountability for the decision, but the risk of error may not be fully visible or accessible to the human actor due to partial opacity related to the evolution of the AI system and the automatic acceptance of the machine-led suggestions without proper checks by the human actor of the selection and categorization of the case law. In light of the AI Act, it can be qualified as a high-risk AI system because of its use in assisting judges in researching prior case laws interpreting specific legal provisions that can be then applied to new sets of facts.[10] Furthermore, there is partial opacity in the system and a risk of error that may not be fully visible because of the self-learning characteristic of the systems. If the system will be put into use to generally support the activity of the judiciary in Germany, the system may require a prior fundamental rights impact assessment in accordance with the AI Act

[7] https://jura.uni-koeln.de/en/fakultaet/zentrale-einrichtungen/studien-und-karriereberatungszentrum/newsletter/april-2020/news-april-2020/rueckblick-des-legal-tech-lab-cologne (accessed on 4 February 2024).

[8] For example, see the potential dangers to procedural fairness when using information provided in cases in which the defendant was not a party: ECtHR, *Khodorkovskiy and Lebedev v. Russia* (no. 2), nos. 5111/07 and 42757/07, 14 January 2020, § 522–523; *Huseyn and Others v. Azerbaijan*, nos. 35485/05 and three others, § 212, 26 July 2011; the European Court of Human Rights, Guide on Article 6 of the European Convention on Human Rights, Right to a Fair Trial (Criminal Limb), Updated on 30 April 2020, p. 25; as well as national procedural rules requesting the evaluation of individual elements of evidence provided by the parties and the public prosecutor in coming to a decision in a case brought before the court (e.g. Article 115 Italian Code of Civil Procedure, Articles 6–9 French Code of Civil Procedure).

[9] Frauke Rostalski, Smart Sentencing, Conference organized by the German Presidency of the European Council, *Digitalisation of Justice—Interconnection and Innovation*, 8 December 2020.

[10] Article 6(2) in conjunction with Point 8 Annex III and Recital 61 in AI Act.

to identify the potential risks for parties, as well as a conformity assessment and a process of periodic control to ensure the good performance and safety of the system while in use.[11]

7. Predictive Systems: The Unbridgeable Accountability Gap

This section explores systems designed to support the judicial decision process offloading part of the decision's complexity from the judge, orienting their choice and, in some instances, helping to justify their decision. Therefore, such tools may intentionally or unintentionally steer judicial decision making, shape sentencing practices, and influence parties' treatment. These kinds of systems were in place before the recent AI diffusion, and some of them are simple tables and guidelines. An example is the Finnish "unofficial guidelines based on court practices and organized judicial discussions" (Hinkkanen and Lappi-Seppälä 2011). In drunk driving cases, these guidelines help the judge to orientate themself in the decision-making process while weighing elements such as alcohol concentration, recidivism, age, and possible previous sentences of the (alleged) offender. The system was designed to help in "adjusting penalties to the degree of risk and hence of presumed culpability in the apprehended behaviour" (Ross et al. 1984). According to the "score" of each variable, the case can be associated with a recommended sentence. The Finnish guidelines are non-binding and are built and updated by judges based on law, court practices analysis, policy objectives, and open group discussion. As the system influences the decision, judges worried that these schematic tables could curtail their judicial discretion in an unacceptable manner. At the same time, the objectives of the influencing process are explicit: to "reduce the use of unconditional imprisonment and disparities in sentencing" (Hinkkanen and Lappi-Seppälä 2011, pp. 378–79). Additionally, the collective reflection process takes place within the judiciary under effective judicial control. The result is increased predictability in judicial decision making and equal treatment.

Looking at technology-enabled predictive systems, COMPAS emerges as a much-discussed example. More precisely, COMPAS is a case management system which includes a risk assessment tool based on 22 modules (risk and needs scales) built around 137 variables. It is used to "inform" decisions based on risk and need factors identified by criminological studies and correctional practices (Equivant 2017, pp. 1–2). When using a risk model, a profile of the offender is generated, combining current offense(s) and criminal history data with data collected through a close-ended questionnaire which can be self-filled or filled in by the probation officers (Northpointe 2012). The system compares the profile generated to calculate the person's risk scores (Michigan Department of Corrections 2017, p. 5). While the risk assessment is used to influence the decision of law enforcement and judicial officers for a single individual case, the tool predicts group behaviors. In other words, the risk score is an estimation based on the known outcomes of groups of offenders who have characteristics similar to the ones of the individual being assessed (Equivant 2019, p. 35). While there is some possibility to tune the system by the authority adopting it, the algorithm itself and the weight of the variables are provided by the private company supplying it. When local norm studies are carried out to set the system, such activities are typically carried out by the supplier and not by the judges making use of the system (Dieterich et al. 2014).

While machine learning and other AI techniques are regularly associated with the discussion of COMPAS (see, for instance, Agudo et al. 2024, and Van Dijck 2022), researchers working for the company which developed the application state argue that the software uses statistical methods as correlations and regression and not machine learning (Jackson and Mendoza 2020). If this is the case, COMPAS may not be categorized as an AI system according to the definition of the EU AI Act, although its use and consequences could have effects similar to those of AI systems that are considered high risk according to the AI Act. The confusion is due to the non-public nature of the algorithm (DeBrusk 2018; Hall and Gill 2017; Rudin et al. 2020) and the reference, in the system's presentations,

[11] Articles 9 and 27 in conjunction with Recitals 57 and 96 and 155 in AI Act.

to the possibility of different approaches to a predictive model (Equivant 2019, p. 13). Confusion is further fueled by authors involved in COMPAS development, suggesting that machine learning could contribute to model recidivism. Jackson and Mendoza underline that the agencies using COMPAS have full access to risk variables, scoring processes, and any relevant information and that the system was developed to create simple, directly interpretable, and transparent risk models (Jackson and Mendoza 2020, p. 8). However, as this disagreement about the system's AI nature itself shows, the system may remain opaque even for researchers not directly involved in its development. Doubt can be cast about the capacity to interpret the system's functioning by non-statistically savvy judges.

RisCanvi is another risk assessment system, commissioned by the Catalan Department of Justice to the University of Barcelona in 2008 and introduced in 2009. It is used to assess recidivism and violence risk, determining parole access in the Catalan prison system, and offers similar functionalities to COMPAS (Digital Future Society 2023, p. 16; Galdon-Clavel et al. 2024). The system is built on a series of studies carried out since 1991 by the Department of Justice on prison recidivism and the 2007 recommendations of a Department of Justice Experts Commission for the introduction of a risk assessment protocol for managing dangerous offenders in prisons and after their release. RisCanvi was developed in three phases: (1) identifying risk factors, (2) training users, and (3) implementing it across 25 Catalan prisons within a year, integrating it into management systems and rehabilitation programs. Since 2010, around 15,000 assessments have been completed, with quality control ensured through the SOS RisCanvi office and regular calibration by a team of validators and through psychometric analysis (Andrés-Pueyo et al. 2018, pp. 256–58).

RisCanvi was designed with the objective of improving individualized predictions of inmate behaviors, including violent crimes, self-harm, and in-prison aggression, and assessing the likelihood of breaches during furloughs or parole. The system was intended to support professionals by providing better information, standardizing decision making, reducing errors, and ensuring transparency. Its regular use was intended to promote best practices in information management and to improve decision making through "technical instruments of empirically proven validity and utility" (Andrés-Pueyo et al. 2018, p. 257). The system evaluates an inmate's risk level by assessing various factors, including criminal history, personal background, social environment, medical history, and psychological state. Initially, the risk factors were identified based on internal data (from 2003 to 2008) corresponding to approximately 600 inmates. Since 2010, the system has gone through three iterations and now includes five types of risks: self-directed violence, violence towards other inmates or staff, violent recidivism, breaching parole, and general recidivism (Digital Future Society 2023, p. 16).

The system has a module for screening, which includes 10 risk items and two risk levels (low and high), and a full version, which includes 43 risk items and three risk levels (low, medium, and high). Both versions include the four grouping variables (age, sex, criminal status, and national origin) that moderate "in different ways the effect of risk factors included in the algorithm" (Andrés-Pueyo et al. 2018, p. 259).

Risk assessment information is gathered by professionals like psychologists, criminologists, and social workers from various sources: administrative records (e.g., evaluations by social workers and health professionals), legal records (e.g., lawyers' input), criminological data, and input from inmates' relatives, as well as interviews with offenders and interactions with prison staff. A key issue is that inmates are often unaware of the assessment or the interviews conducted for it (Galdon-Clavel et al. 2024, pp. 17–22).

Judicial records on the crimes committed and correctional details like prison time and sanctions are automatically integrated into RisCanvi and regularly updated before staff conduct evaluations (Andrés-Pueyo et al. 2018, p. 261). The risk is assessed every six months to reflect any changes in the individual's circumstances or behavior (Galdon-Clavel et al. 2024, p. 20). Even if, as with COMPAS, the system is sometimes described as AI (Agudo et al. 2024), risks are calculated through classical statistics like correlations and

regressions. The Prison Administration considered the introduction of machine learning, but experimentation did not show significant improvements (Bellio 2021).

In the case of non-violent crimes, the assessment begins with screening and if the resulting risk is high, it includes the complete test. In the case of violent crimes, a complete assessment is always used. Users can override the rating but must provide justification, while this is not needed if the algorithmic outcome is accepted. Since the calculation behind the risk rating is not clear to professionals, it is difficult for them to discuss and change the rating. As a result, changes in the risk assessment take place in less than 5% of cases, suggesting a significant influence of the algorithm on the final decision (Galdon-Clavel et al. 2024, p. 21).

As the Eticas Adversarial Audit shows, while the 43 risk indicators are known, the actual risk calculation algorithms and the different weights are not provided and cannot be easily calculated based on the available data. The system remains opaque to external scrutiny. Furthermore, as far as the inmates are concerned, the problem is not just the opacity of the algorithm, but the fact that they are not even informed of their risk level and therefore do not meet the conditions to improve their situations, especially considering that over 60% of the risk factors are dynamic (Galdon-Clavel et al. 2024, pp. 21–22) [p. 19] and can therefore be improved (e.g., peer group influence, etc.).

As the confusion in the actual nature and functioning of COMPAS beyond the experts working for Northpointe and the opaqueness of RisCanvi for the operators and inmates show, external actors are not in a position to assess the calculated risk. At the same time, the risk assessment of recidivism, absconding, or violence is not deterministic. Thus, a different system with a different algorithm but a similar percentage of correctness could provide a different assessment on the risk posed by an individual as different factors are considered or differently weighted. Even if the systems are not AI-based, they are non-accountable to users.

With the use of these systems, part of the assessment activities which the judge or the professionals are supposed to carry out are delegated to COMPAS or RisCanvi. As the machine provides an assessment of the risk, the human actor deciding on the case will be influenced by such assessment. The agency in the risk assessment is therefore shifted from the human to the non-human being, but as the system is non-accountable, the human actor remains accountable for the decision. As RisCanvi shows, after being taught that the system may provide counterintuitive results but that such results are based on a solid theoretical background, the user does not appear to be in the position to easily discount its suggestion. The user cannot verify the system's outputs based on its inputs in the way it can be carried out with speech-to-text tools. Organizational arrangements are not built to increase the level of control of judges on or against the traditional tables and guidelines, such as the Finish sentencing guidelines. The judge can only trust (or not trust) the system. But if the judge decides not to trust the system, what will happen if, once released, the defendant commits a crime? Or, in a different scenario, the judge may decide to keep the defendant in custody due to their assessment of the situation even if the system predicts a low recidivism risk. Would the judge be ready to ignore the suggestion of the machine and keep the defendant in custody? As we have seen in the case of RisCanvi, while the human in the loop is required, cases in which the human actors go against the suggestion of the machine are extremely rare in real-world situations (Agudo et al. 2024; Galdon-Clavel et al. 2024, p. 21). While potentially improving the quality of the assessment, the use of such systems has a new influence on judicial decision making.

The human actor remains formally accountable, but the possibility to verify the system's performance remains at the agency level or even at the software provider level, and it is carried out at group level rather than at the individual case level. Judges are accountable for decisions strongly suggested (and de facto taken) by non-accountable systems. All of these contribute to endangering judicial independence and fair trial.

Furthermore, while in systems such as COMPAS and RisCanvi, the algorithm can be assessed by humans (if not concealed for commercial or other reasons), as machine learning

or other AI techniques are introduced, this assessment becomes increasingly difficult even for the developers themselves.

In practice, AI systems make statistical correlations between data in a manner opaque to human understanding. Then, such information is packed, summarized, and delivered to the human judge to "support" their decisional duties. The decision remains with the judge, but it can be hard for the judge to resist such "disinterested" and "science-based" suggestions. Therefore, the risk is that while system developers intend to delegate the activity of proposing a suggestion to the system, they end up achieving the delegation of the decision itself to the system.

8. In Search of Sound Accountability Mechanisms

When activities traditionally performed by humans are delegated to technology, new artifacts enter the actor–network configuration that handles judicial proceedings. This transition brings about a range of changes. If new technologies are accountable (i.e., it is possible to assess the correctness of the output or of the process), it becomes possible to evaluate whether they meet design requirements and to differentiate between human and machine accountability. This normally works with traditional (non-AI-based) digital technologies used by courts since the 1990s. On the contrary, with non-accountable AI systems, agency is delegated to technology, while accountability remains entirely with the users. In the actor–network configuration, agency and accountability are decoupled. This has been observed in speech to text, in the Smart Sentencing project, and in the case of predictive decisions. Users become the sole being responsible for the consequences of using technological devices based on algorithms that, for whatever reasons, remain obscure or not (completely) under their control. It is then critical to ascertain in which cases (1) the decoupling of agency and accountability is acceptable in judicial proceedings and (2) how it interferes with the complex sets of values underpinning the judicial function.

The focus on the accountability of the entire actor–network configuration—and not just on technology—eases the assessment and the identification of viable solutions. As noticed, in speech-to-text applications, the judge can easily check the level of matching between the spoken act and the written text. They have all the information and knowledge to check the level of matching and assume the responsibility for the deed. AI-based algorithms do not disturb the tasks of the judges in this case as they can easily verify the correctness of the outputs generated by the technology by confronting them with the input. This check can be introduced in working practices without increasing workload. Judges can remain accountable for the contents of the deeds if the text is written with a pencil, a keyboard, or speech-to-text software. In this case, the decoupling between agency and accountability and the changes (improvements) in the AI algorithm is not an issue. *The first conclusion is that using obscure and non-accountable technologies is not an issue per se*.

Legal analytics cases are more complex. Sentenze Web and Smart Sentencing, the two examples discussed, show systems that select and organize data along pre-established criteria, allow users to extract and compare judgements, or check their own legal issue against pre-existing decisions. Databases are not designed to execute discrete tasks, such as speech-to-text, but to analyze (and, to some extent, influence) legal and judicial affairs by the outcomes they provide to the user. Sentenze Web and Smart Sentencing perform similar functions of extracting relevant data from a body of sentences to enable various kinds of legal, jurisprudential, or factual analysis. However, this is carried out through two different actor–network configurations in which the roles of humans and machines are combined in various ways. In the case of Sentenze Web, the creator of the system—the same Court of Cassation—classifies the cases and makes clear the criteria adopted and the functioning of the search engine. The search results are limited to the selection of sentences collected in the corpus juris. The software code is stable, and it can change just with human intervention. The features of this actor–network configuration allow for clear accountability. The agency assigned to the system is carried out to make available and select relevant jurisprudence. The system's owners are accountable for its functioning; the users are accountable for

the actions and decisions taken or influenced by the information provided by the system. If the system introduces biases, for instance, excluding sentences with a jurisprudential orientation that diverges from the mainstream, or it is poorly classified, the system's owner can be made accountable for the deficiencies. The transfer of agency (from a search made in the library to a search made with Sentenze Web) is coupled with a symmetrical transfer of accountability.

In the case of Smart Sentencing, the different configuration of the actor–network configuration makes the creation of effective accountability arrangements critical. The classification of sentences is carried out using a mix of AI-based algorithms (unsupervised machine learning, neural network, and natural text/speech recognition) and human inputs at the stage of supervised machine learning and when judges check the classifications suggested by the algorithm. Such checks provide feedback to the machines and result in automatic changes in the algorithm itself. The actor–network architecture assigns to the judge the final check and the quality of the data entered into the database. Hence, despite the amount of work conducted by the technology to extract information from the texts of the sentences, the individual judges are accountable for the quality of the entry (suggested by the machine but confirmed by each judge issuing the decision). Besides this, the individual judge has no understanding of the inner functioning of the system and, thus, of the mechanisms through which sentences are extracted and classified by the system upon queries. At the same time, the final user is accountable for the system's use at the decision-making level. As in speech-to-text technology, there is decoupling between the agency delegated to the machine and the accountability that remains attached to the users. Consequently, judges are accountable for both the inputs (i.e., the data entered) and the outputs of the system (i.e., the decisions made considering the analysis carried out by the machine). But if the judge using speech-to-text technology is in total control of the input–output links, is this the case with judgment databases? Judges are fully equipped to categorize judicial decisions accurately and use search engines like Sentenze Web to retrieve cases and relevant data. In the actor–network configuration, the judge's accountability can be easily distinguished from that of the system. Smart Sentencing shows a different configuration. When the judges have no time (or attention) to carefully check the classification suggested by the machine, hence limiting themselves to take over and accept the algorithm's suggestion, a risk of error sneaks in. The users' passive attitude can be resolved by promoting the collective use of the systems by court sections or departments to analyze jurisprudence and conduct collective checks of the classifications provided by different judges. This could also allow for collective control of the relations between inputs and outputs and is a condition to be ascertained in order to hold the judges accountable for the use of the system. Furthermore, differently from Sentenze Web, the AI-based algorithms used to extract sentences are dynamic and evolve over time. As the concept of "machine learning" suggests, software codes evolve through use and with the upload of new sentences to train the system. Therefore, judges—with their legal skills and institutional position—can be held accountable just for the initial semantic annotation of the sentences and not for the day-by-day evolution of the algorithm and, hence, for the outcome of the system. However, they will be accountable for their decisions, influenced by the outcome of smart sentencing. Agency is transferred to the machine, but accountability remains attached to the final users, who are not in the condition to easily check the outputs to identify machine-based issues. To face this problem, a third party should intervene to ensure the lack of bias in nudging towards certain judicial orientations or other undue influences on judicial decision making. Following the AI Act, this should be certified by an assessment of the system before its introduction and during its life cycle.[12] The Act's application will reinforce requirements for reducing risks, maintaining human oversight, and correcting actions that affect the proper functioning of the application when AI systems are used by courts to support the interpretation of the

[12] See Articles 6, 8, and 9 in AI Act.

law and the decision-making process.[13] Additionally, with a view to eliminate or reduce the risks related to high-risk AI systems, the systems' users—the judges—should be given appropriate training to acquire technical knowledge and experience with systems such as Smart Sentencing.[14] All of this injects complexity and costs, making the use of these systems more onerous and less efficient. Hence, accountability requires the establishment of complex organizational arrangements in which the work of judges to check data quality must be supplemented by third parties' contributions and certifications. In this second case, *accountability has economical and organizational costs that must be justified by the benefits of the new technological arrangement*.

Looking at predictive systems, such as COMPAS and RisCanvi, agency is transferred to the technological system and, as in Smart Sentencing or speech-to-text technology, the judge remains accountable for the use of the system's outcomes. Are the judges then capable of understanding the input–output relations that entail how data are transformed into a score signaling the recidivism risk? Are they capable of properly considering the probabilistic nature of the assessment based on group data (i.e., false positive and false negative results)? The question is more acute here because the agency transferred to technological artifacts entails the evaluation of group factors deemed to be relevant in assessing concrete individual cases retained to fit within the pre-established group. Regarding predictive systems, it is the core of the judicial function that has to be protected from external influences.

As with speech-to-text technology and Smart Sentencing, the way in which the inputs are transformed into outputs in COMPAS and RisCanvi remains obscure for specialists and judges. Although the judge can control the outputs of speech-to-text technology and supervise semantic annotations in Smart Sentencing (or delegate quality assurance of the system to a third party), such mechanisms are not in place in COMPAS or RisCanvi. Judges and other professionals may know the inputs (i.e., the answers to the questionnaire of the accused and other data), but they cannot directly relate them to the outputs. While formally accountable for the decision taken, the single judge (or the professional validating the outputs) does not have the means to accurately assess the "suggestion" the machine makes.

Mechanisms could be developed to improve the accountability of the technological components of the systems (Diakopoulos 2016). In the case of AI systems aimed at *supporting judicial decision making*, the process will be reinforced by the provisions of the EU AI Act applicable starting in the second half of 2026. This includes addressing the system's opacity and allowing external checks. The Act sets specific duties for assessment, testing, registration, and access to AI models both ex-ante and ex-post, particularly for those classified as high-risk AI systems used in justice services and enforcement (European Commission 2021, chap. 2). Such actions aim to prevent and limit threats and undesired manipulations that the models may introduce.[15] At the same time, the solution proposed by the EU AI Act of ex-ante and ongoing external check does not provide the judge with the means to ascertain that the suggestion provided by the machine does not result in prohibited practices such as "detrimental or unfavorable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behavior or its gravity" (Art 5 EU AI Act). This check is delegated to third parties and on an aggregated level (i.e., not based on individual cases). Considering the kind of influence the system aims to provide to deliberations, the judge has to trust third parties, and their direct oversight can only be a limited one. In this regard, "human oversight is in danger of becoming a value in itself, an empty procedural shell used as a stand-in justification for algorithmization but failing to provide protection for fundamental rights" (Koulu 2020).

Even with the safeguards of the AI Act, judges will be in the awkward position of receiving suggestions from a machine that remains non-accountable, while they are ac-

[13] See Articles 14, 15, and 20 in AI Act.

[14] See Art 9(5) in AI Act.

[15] See Section 3.5 Fundamental Rights, Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM (2021) 206 final.

countable for the decision made with the machine's contribution, but without being able to properly assess the correctness of the suggestion. Furthermore, unless specifically indicated, lawyers, prosecutors, and parties will assess the judicial decision without having the possibility to understand to what extent the judge has decided considering the suggestions of the machine and/or the reasoning behind the suggestion made by the system. In this regard, the new configuration of the actor–network configurationseems incompatible with the principles of judicial independence and fairness enshrined in Article 6(1) ECHR. All the risks of potential undue judicial influences passing through the peephole of non-accountable decision support systems occur with COMPAS and RisCanvi. As discussed, it is not clear whether the two systems are AI systems. Hence, it is unclear whether they will require the scrutiny established by the AI Act. *This shows how this issue emerges not just with AI systems, but with any non-accountable system influencing judicial decision making.*

Moving the assessment of accountability from the technological artifact to the actor–network configuration of humans and technological agents leads to the identification of dynamics that affect the digital innovation in the administration of justice while having a more comprehensive view of the impact of AI. This shows a paradox: totally non-accountable AI systems can be adopted as speech-to-text systems without harm if the user who becomes accountable for the output of the system can exercise reasonable control over their outputs. Other non-accountable systems, such as COMPAS and RisCanvi, can become serious sources of undue influence and menaces to judicial independence regardless of whether they are or not AI-based. Hence, the question is not just related to AI, but rather on the introduction, into judicial proceedings, of non-accountable technologies that cannot be supervised, controlled, or understood by judges, lawyers, and case parties.

Author Contributions: Conceptualization, methodology, data collection, analysis, and writing are the result of a collective effort of the three authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study does not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

Agudo, Ujué, Karlos G. Liberal, Miren Arrese, and Helena Matute. 2024. The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications* 9: 1–16. [CrossRef] [PubMed]

Andrés-Pueyo, Antonio, Karin Arbach-Lucioni, and Santiago Redondo. 2018. The RisCanvi. In *Handbook of Recidivism Risk/Needs Assessment Tools*. Oxford: John Wiley & Sons, pp. 255–68.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchne. 2016. Machine Bias. There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica*. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 14 November 2024).

Ashley, Kevin D. 2017. *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*. Cambridge: Cambridge University Press.

Association for Computing Machinery and US Public Policy Council. 2017. *Statement on Algorithmic Transparency and Accountability*. Washington, DC: Association for Computing Machinery and US Public Policy Council.

Atchison, Amy B., Lawrence Tobe Liebert, and Debuse K. Russell. 1999. Judicial Independence and Judicial Accountability: A selected bibliography. *Southern California Law Review* 72: 723–810.

Bellio, Naiara. 2021. In Catalonia, the RisCanvi Algorithm Helps Decide Whether Inmates Are Paroled. algorithmwatch.org. Available online: https://algorithmwatch.org/en/riscanvi/ (accessed on 14 November 2024).

Bijker, Wiebe E., and John Law, eds. 1992. *Shaping Technology Building Society. Studies in Sociotechnical Change*. Cambridge, MA: The MIT Press.

CCJE. 2023. *Compilation of Responses to the Questionnaire for the Preparation of the CCJE Opinion No. 26 (2023) "Moving Forward: Use of Modern Technologies in the Judiciary"*. Strasbourg: Council of Europe.

Chiao, Vincent. 2019. Fairness, accountability and transparency: Notes on algorithmic decision-making in criminal justice. *International Journal of Law in Context* 14: 126–39. [CrossRef]
Contini, Francesco. 2020. Artificial Intelligence and the Transformation of Humans, Law and Technology Interactions in Judicial Proceedings. *Law, Technology and Humans* 2: 4. [CrossRef]
Contini, Francesco, and Giovan Francesco Lanzara. 2014. *The Circulation of Agency in E-Justice. Interoperability and Infrastructures for European Transborder Judicial Proceedings*. Berlin/Heidelberg: Springer.
Czarniawska, Barbara. 2004. On time, space, and action nets. *Organization Studies* 11: 773–91.
Czarniawska, Barbara, and Bernward Joerges. 1998. The Question of Technology, or How Organizations Inscribe the World. *Organisation Studies* 19: 363–85.
DeBrusk, Chris. 2018. The Risk of Machine-Learning Bias (and How to Prevent It). *MIT Sloan Management Review*. Available online: https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/ (accessed on 14 November 2024).
Dhungel, Anna-Katharina, and Eva Beute. 2024. AI Systems in the Judiciary: Amicus Curiae? Interviews with Judges on Acceptance and Potential Use of Intelligent Algorithms. Paper presented at ECIS 2024, Paphos, Cyprus, June 13–19.
Diakopoulos, Nicholas. 2016. Accountability in Algorithmic Decision Making. *Communications of the ACM* 59: 56–62. [CrossRef]
Dieterich, William, William L. Oliver, and Tim Brennan. 2014. *COMPAS Core Norms for Community Corrections*. Northpoint. 97. Available online: https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-WIDOC_DCC_norm_report021114.pdf (accessed on 14 November 2024).
Digital Future Society. 2023. *Algorithms in the Public Sector: Four Case Studies of ADMS in Spain*. Barcelona: Digital Future Society.
ENCJ. 2018. *Independence, Accountability and Quality of the Judiciary*. Adopted General Assembly Lisbon, 1 June 2018. European Network of Councils for the Judiciary. Bruxelles: ENCJ. Available online: https://pgwrk-websitemedia.s3.eu-west-1.amazonaws.com/production/pwk-web-encj2017-p/Reports/ENCJ%20Report%20IAQ%202017-2018%20adopted%20GA%20Lisbon%201%20June%202018.pdf (accessed on 14 November 2024).
Equivant. 2017. *Northpointe Specialty Courts Manage Your Treatment Docket*. Northpoint. Available online: http://www.equivant.com/wp-content/uploads/Northpointe_Specialty_Courts.pdf (accessed on 14 November 2024).
Equivant. 2019. *Practitioner's Guide to COMPAS Core*. Northpoint. Available online: https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf (accessed on 14 November 2024).
European Commission. 2021. *Proposal for a AI ACT. Explanatory Memorandum*. Brussels: European Commission.
European Parliament. 2024. *EU AI Act: First Regulation on Artificial Intelligence*. Bruxelles: European Paarliament, Directorate General for Communication.
Galdon-Clavel, Gemma, Mat Mastracci, Luis Rodrigo González Vizuet, and Miguel Azores. 2024. *Automated (In)Justice? An Adversarial Audit of RisCanvi*. Barcellona: Eticas Foundation. Available online: https://eticasfoundation.org/?audit-spotlight=the-adversarial-audit-of-riscanvi (accessed on 14 November 2024).
Galli, Federico, and Giovanni Sartor. 2023. AI approaches to predictive justice: A critical assessment. *Humanities and Rights Global Network Journal* 5: 165–217.
Garapon, Antoine, and Jean Lassègue. 2018. *Justice Digitale: Révolution Graphique et Rupture Anthropologique*. Paris: Presses Universitaires de France.
Garapon, Antoine, and Jean Lassègue. 2021. *La Giustizia Digitale. Determinismo Tecnologico e Libertà*. Original edition: Justice digitale. Révolution graphique et rupture anthropologique, Paris, Presses Universitaires de France, 2018 ed. Bologna: Il Mulino.
Goodman, Bryce, and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38: 50–57. [CrossRef]
Hall, Patrick, and Navdeep Gill. 2017. Debugging the Black-Box COMPAS Risk Assessment Instrument to Diagnose and Remediate Bias. *Open Review*. 6. Available online: https://api.semanticscholar.org/CorpusID:6578316 (accessed on 14 November 2024).
Hinkkanen, Ville, and Tapio Lappi-Seppälä. 2011. Sentencing Theory, Policy, and Research in the Nordic Countries. *Crime and Justice* 40: 349–404. [CrossRef]
Hu, Xia, and Huam Liu. 2012. Text Analytics in Social Media. In *Mining Text Data*. Edited by Charu C. Aggarwal and Cheng Xiang Zhai. Berlin/Heidelberg: Springer, p. 388.
Jackson, Eugenie, and Christina Mendoza. 2020. Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not. *Harvard Data Science Review* 2: 2644–353. [CrossRef]
Johnson, Deborah G. 2004. Computer Ethics. In *The Blackwell Guide to the Philosophy of Computing and Information*. Edited by Luciano Floridi. Oxford: Blackwell, pp. 64–74.
Kirsch, Birgit, Sven Giesselbach, Timothée Schmude, Malte Völkening, Frauke Rostalski, and Stefan Rüping. 2020. Using Probabilistic Soft Logic to Improve Information Extraction in the Legal Domain. Paper presented at the LWDA, Online conferennce, September 9–11.
Koulu, Riikka. 2020. Human Control over Automation: EU Policy and AI Ethics. *European Journal of Legal Studies* 12: 9–46.
Kyriakides, Nicolas, Anna Plevri, and Yomna Zentani. 2022. AI and access to justice: An expansion of Adrian Zuckermans findings. In *Frontiers in Civil Justice*. Edited by Xandra Kramer, Jos Hoevenaars, Betül Kas and Erlis Themeli. Cheltenham: Elgar, pp. 121–41.

Lanzara, Giovan Francesco. 2009. Building digital institutions: ICT and the rise of assemblages in government. In *ICT and Innovation in the Public Sector. European Studies in the Making of E-Government*. Edited by Francesco Contini and Giovan Francesco Lanzara. Basingstoke: Palgrave Mcmillan, pp. 9–49.
Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
Lessig, Lawrence. 2007. *Code and Other Laws of Cyberspace. Version 2.0*. New York: Basic Books.
McGregor, Lorna, Daragh Murray, and Vivian Ng. 2019. International Human Rights Law as a Framework for Algorithmic Accountability. *International & Comparative Law Quarterly* 68: 309–43. [CrossRef]
Michigan Department of Corrections. 2017. *Administration and Use of COMPAS in the Presentence. Investigation Report*. Lansing: State of Michigan, Department of Corrections.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3: 1–21. [CrossRef]
Monteiro, E. 2000. Actor-Network Theory. In *From Control to Drift*. Edited by Claudio Ciborra. Oxford: Oxford University Press.
Nissenbaum, Helen. 1996. Accountability in a computerized society. *Science and Engineering Ethics* 2: 25–42. [CrossRef]
Nissenbaum, Helen. 2007. Computing and Accountability. In *Computer Ethics*. Edited by John Weckert. London: Routledge, pp. 273–80.
Northpointe. 2012. *COMPAS Risk & Need Assessment System*. Lansing: State of Michigan, Department of Corrections.
Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. 2024. Accountability in artificial intelligence: What it is and how it works. *AI & Society* 39: 1871–82. [CrossRef]
Pégny, Maël, Eva Thelisson, and Issam Ibnouhsein. 2019. The Right to an Explanation. An Interpretation and Defence. *Delpi* 4: 161–66.
Pierce, Natalie, and Stephanie Goutos. 2024. Why Lawyers Must Responsibly Embrace Generative AI. *Berkeley Business Law Journal* 21: 51. [CrossRef]
Reiling, Dory. 2020. Courts and Artificial Intelligence. *International Journal For Court Administration* 11: 1–10. [CrossRef]
Ross, H. L., H. Klette, and R. McCleary. 1984. Liberalization and rationalization of drunk-driving laws in Scandinavia. *Accident Analysis & Prevention* 16: 471–87.
Rostalski, Frauke, Timothée Schmude, Malte Völkening, and Jin Ye. 2021. Smart Sentencing Grundriss einer teilautomatisierten Strafzumessungsdatenbank. *LRZ*, 166–78. Available online: https://lrz.legal/images/pdf/SmartSentencing.pdf (accessed on 14 November 2024).
Rudin, Cynthia, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review* 2. [CrossRef]
Seibert-Fohr, Anja, ed. 2012. *Judicial Independence in Transition*. Heidelberg: Springer.
Shah, Hetan. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A* 376: 20170362. Available online: https://royalsocietypublishing.org/doi/10.1098/rsta.2017.0362 (accessed on 14 November 2024). [CrossRef] [PubMed]
Simon, Herbert A., Donald W. Smithburg, and Victor A. Thomson. 1961. *Public Administration*, 6th printing (f. p. 1950) ed. New York: Alfred A. Knopf.
Spasojevic, Dijana, Miglena Vucheva, Margarida Rocha, Robrecht Renard, and Dimitrios Stasinopplous. 2020. *Study on the Use of Innovative Technologies in the Justice Field—Final Report*. Directorate-General for Justice and Consumers (European Commission). Bruxelles: European Commision Publication Office.
Van Dijck, Gijs. 2022. Predicting Recidivism Risk Meets AI Act. *European Journal on Criminal Policy and Research* 28: 407–23. [CrossRef]
Velicogna, Marco. 2007. Justice Systems and ICT: What Can Be Learned From Europe? *Utrecht Law Review* 3: 129–47. [CrossRef]
Velicogna, Marco. 2023. A Time for Justice? Reflecting on the Many Facets of Time and Temporality in Justice Service Provision. In *Organization as Time. Technology, Power and Politics*. Edited by François-Xavier de Vaujany, Robin Holt and Albane Grandazzi. Cambridge: Cambridge University Press, pp. 349–74.
Zuckerman, Adrian. 2020. Artificial Intelligence—Implications for the Legal Profession. *Adversarial Process and Rule of Law Law Review Quarterly* 136: 427–53.