UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the journal article:

Rovira Martorell, Joan; Gálvez, Ana; Tirado Serrano, Francisco Javier. «Artificial intelligence and the production of judicial truth». Theory Culture and Society, 2024. DOI https://doi-org.are.uab.cat/10.1177/0263276424126817

This version is available at https://ddd.uab.cat/record/303301

under the terms of the IN COPYRIGHT license

Check for updates

Article

Theory Culture & Society

Theory, Culture & Society
1–16
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02632764241268174
journals.sagepub.com/home/tcs
Sage

Artificial Intelligence and the Production of Judicial Truth

Joan Rovira Martorell
Universitat Autònoma de Barcelona

Ana Gálvez
Universitat Oberta de Catalunya

Francisco Tirado
Universitat Autònoma de Barcelona

Abstract

The aim of this paper is to present artificial intelligence (AI) as an organ with a role in the production of judicial truth, expanding its objects, changing its procedures and reshaping the distribution of agencies within the judicial organism. To this end, it builds on Michel Foucault's work on the procedures of truth production and the three subject forms involved: operator, spectator and object. This is then complemented by the general organological perspective proposed by Bernard Stiegler. On the basis of both, we will demonstrate two realities: first, that AI is shifting truth production from the individual to the profile, and second, that the types of associations that AI is forming have the potential to curtail human agency in the production of judicial truth.

Keywords

artificial intelligence, justice, profile, risk, truth production

Introduction

In recent years, artificial intelligence has been introduced into judicial processes, serving mainly two distinct functions: *policing*, with tools such as PredPol,[1] for which AI has been used in police institutions to statistically predict the geographical areas where crime is most likely to occur and thus optimise the deployment of police patrols; and *sentencing*, where it supports the various phases of the judicial process, including choosing and assigning charges (tools such as Víctor[2] and PretorIA[3]), advising judges on the type of

Corresponding author: Francisco Tirado. Email: FranciscoJavier.Tirado@uab.cat
TCS Online Forum: https://www.theoryculturesociety.org/

prison sentence to hand down based on the offender's risk profile (e.g. HART[4] and COMPAS[5]), automatically handling petty crimes (e.g. Prometea[6]), and predicting court rulings so that law firms can develop strategies accordingly (e.g. Lex Machina[7]).

Sentencing AIs are the ones that interest us in this paper, as our aim is to show their agency in the judicial manifestation of truth. They are also the most controversial because of the dangers and risks they pose, and they are the focus of most reflections on the introduction of artificial intelligence in judicial processes. *Wisconsin v. Loomis* is often presented as a paradigmatic example of these dangers (Casabona, 2018; Ulenaers, 2020). Eric Loomis was sentenced to six years in prison in July 2016, this being the first time that the AI-driven software COMPAS had been used to dictate a prison term. The defence appealed the sentence on three grounds: (1) that the defence could not be made aware of the AI procedure, as its code was protected as a trade secret; (2) that Loomis' right to an individualised sentence had been denied, as the AI uses group data; and (3) that he had been discriminated against on the basis of gender, as he had a higher associated risk of recidivism as a male. The appeal was dismissed, creating the first precedent of a ruling involving an AI.

The three arguments of Loomis' defence embody the three main concerns about the use of AI in judicial processes. Regarding the first argument, which has to do with the transparency of algorithmic processes, researchers have pointed out the need to know what personal and legal data the AI has been fed, e.g. what historical periods they are from or how legal inconsistencies are processed (Battelli, 2020); the need to avoid black box data processing (Ulenaers, 2020); and the need for the coding of these AIs to be open and accessible (Rizer and Watney, 2018). With respect to the second argument, Casabona (2018) explained that distrust towards the introduction of AI in the judicial system goes back to a long-standing debate between the perspective of criminal dangerousness (subjective evaluations) and that of group risk (statistical techniques) in criminal profiling. The problem with the statistical perspective is that group behavioural patterns are taken as individual characteristics over which the accused has no power to act. The third argument points to the danger of these tools perpetrating and exacerbating existing social biases and discriminations. The issue is that AIs use data on sensitive variables, such as gender and race, because they present high correlations with crime rates, without delving into their real causes (Battelli, 2020; Casabona, 2018; Ulenaers, 2020). The big debate around this issue, as some authors have noted (Berk et al., 2021; Bex and Prakken, 2021), stems from a desire to know to what extent these variables can be omitted without causing the model to become too imprecise.

There is, however, broad consensus that AIs are much more efficient at judicial tasks that do not require much subjective judgement, such as choosing and assigning court charges to cases (Battelli, 2020; Jin and He, 2020) or even autonomously handling petty crimes and setting fines (Levmore and Fagan, 2019; Ulenaers, 2020), which helps to reduce waiting times and diminish their adverse effects on the public.

All of this is traversed by an underlying debate: the link between the rise of AI and the production of truth in judicial processes. The discussion is premised on the idea that AI could embody the possibility of creating truth without the need to develop hypotheses or theorise concepts. It would operate by drawing truth straight out of correlations in the data used to train the algorithmic model (Rouvroy and Berns, 2013). More extreme

positions in this regard claim that AI renders scientific theory obsolete (Anderson, 2008) or that the notion of truth no longer applies insofar as human agents are being replaced and the future will be completely governed by automatisms, annihilating all possibility of contingency (Sadin, 2020).

This paper is framed in this debate and aims to offer a description and model of how AI is involved in the production of truths in the judicial institution. To this end, two dimensions of judicial AI activity are described with empirical examples. Firstly, we show what new forms of knowledge emerge from AI operations, shifting the focus from the individual to the profile. Secondly, we analyse how AI-related truth and knowledge relate to the human capacity for agency and volition. The contribution of this text to the above debate is therefore twofold. On the one hand, it shows how the use of AI in the judicial system produces three different types of truth, the main effect of which is to shift truth production from the individual to the profile. On the other hand, it describes the types of associations formed by AI that have the potential to curtail human agency in the production of judicial truth. While this text has a clear theoretical format, it is based on research that we have been conducting for the last two years, examining the socio-cultural effects of AI in various spheres of everyday life.

To this end, we will first review Foucault's work to conceptualise the forms of truth production and their relationship to the exercise of power, the governance of behaviour and the production of subjectivities. We will then use Stiegler's general organology, and more specifically his concepts of exorganism and pharmacology, to conceptualise the relationships between humans and technology in the judicial institution, as well as the distribution of agencies in its governance. Afterwards, we will explore the changes in the truth produced by the judicial institution, describing how AI participates in judicial truth manifestation in the three forms described by Foucault (2014a): as operator, as spectator and as object. Finally, drawing on the analyses presented, we will offer some considerations on how best to interact with these tools in order to prevent human agency from being displaced by automatisms.

Truth and Judicial Forms

In his later work, Michel Foucault posed questions about truth-telling in relation to the binomials of knowledge-power and subject-government (Foucault, 1979, 2014a, 2014b). Truths, he said, must follow two rules in order to be effective: first, they must be situated within certain orders of discourse, that is, they must adhere to the discursive principles and norms that allow a truth to be stated (Foucault, 2010). Second, they must be faithful to a certain alethurgy, that is, the 'set of procedures, verbal or otherwise, by which one brings to light what is posited as true, as opposed to the false' (Foucault, 2014b: 24). Alethurgy and orders of discourse constitute, in any given historical and cultural context, what Foucault called regimes of truth. Regimes of truth always imply certain obligations on subjects with regard to the manifestation of truth, following a certain rule and logic. They also imply an individual who accepts truths precisely because they are self-evident and logical, in contrast to the mad subject who is incapable of distinguishing truth from falsehood.

Foucault (1996) provided an in-depth analysis of the link between truth and power in the judicial institution on the basis of the historical evolution of the different truth regimes. His work shows that ancient Greek and Germanic law formed the basis of the feudal law that emerged in Europe in the Middle Ages. In the feudal judicial system, breaking the law was an attack on the sovereign's own body. For that reason, when a crime was alleged to have been committed, a representative of the sovereign – the procurator – was dispatched to perform two tasks: first, to conduct an investigation and gather evidence in the form of oral testimony to prove that the alleged offender had indeed committed the crime (ancient Greek law); and second, if found guilty, to administer severe bodily punishment so that the offence against the sovereign could be atoned for (Germanic law). The investigation sought above all one kind of definitive proof: a confession. This, according to Foucault (2014a, 2014b), is an act of affirmation in which the individual commits himself to the stated truth, subjectivises himself in it and embodies it. The act of confession was essential in bringing the truth to light and concluding the judicial or inquisitorial process.

Confession therefore establishes a first link between subjectivity and truth (the individual accepts his criminal status) and another with power (through confession the divine or sovereign order is restored). Confession is thus the materialisation or the very act in which the notions of subjectivity, truth and power are knotted together in a single meaningful totality. These forms of justice changed in the 18th century with the rise of capitalism. Crime was no longer an attack on divine or sovereign law, but on society and its orderly functioning. From then on, judicial truth would no longer seek to restore the power of the sovereign or God, but to right a social wrong. This would be enforced by solitary confinement in prison. As a result, by the 19th century, the exercise of manifesting judicial truth was distributed among a series of prison-like institutions (schools, hospitals, mental asylums, barracks, etc.), which were responsible for monitoring and gathering information about individuals in order to make their bodies and souls productive for the state. It was no longer a question of investigating and searching for evidence, but of examining behaviour in order to steer it in the right direction. Confession ceased to be an act brought about by torture and instead became a constant and ongoing relationship between the individual and his teacher, priest, foreman, guard, psychologist, etc. (Foucault, 2014a, 2014b). This meant that the knowledge produced by these institutions would now feed into the judicial investigation, which sought not only to determine guilt but also to understand the criminal's motive. The criminal went from being someone who had broken the social pact to being a failure of the social mechanism of normalisation, that is, the criminal became the anomalous subject (Foucault, 2003).

Foucault's work thus establishes a close and direct link between power, knowledge and subjection. This involves different regimes for manifesting an individual's truth. Thus, sovereign power rules through law and torture, and knowledge is produced by investigating the guilt of the individual, who ultimately becomes a (guilty) subject by confessing. Disciplinary power, on the other hand, governs through norms: there are institutions of confinement that produce knowledge about each individual, determining whether the subject is normal or abnormal, in order to take corrective measures to prevent him from becoming a criminal.

Risk Management and Truth

The 20th century displaced disciplinary institutions and built-up security apparatuses (Foucault, 2009, 2010). The focus turned to managing natural laws, not to change them, but to reduce the risk of social disasters such as epidemics, famine or high crime rates. Their aim was not so much to normalise the individual as to act on the population, and their main tool was statistics. Risk was defined not as an event that actually occurs, but as a specific way of dealing with certain events that may virtually occur in a population (Ewald, 1991). In this sense, insurers became the privileged institutions in the management of the risk market. According to them, risk has three main characteristics: (a) it is calculable; (b) it is collective in that it does not attribute responsibility to the individual, because the calculation is not related to an act, will or negligence, but to a ratio associated per se with the individual; and (c) it is capital, i.e. a monetary value associated with the calculation of individual risk. In the judicial institution, risk management is embodied in actuarial methods, which are a statistical way of determining an individual's risk of offending in relation to one or more group traits (Harcourt, 2007).

Actuarial methods emerged in the United States in the 1930s as a way of estimating an offender's risk of reoffending. A statistical approach was developed to correlate certain population factors with the risk of individual recidivism (e.g. previous convictions, drug use, level of education, type of employment). Each of these factors was given a score according to its importance, and the judge was responsible for creating a profile of the offender's dangerousness based on the number of factors he met. The higher the risk, the harsher the sentence. The introduction of AI has increased the capacity of these methods to process huge amounts of data, to personalise the weight of risk factors according to the context of each individual, and to expand the objects on which judicial truth is produced. However, since its inception, risk management has shifted its focus from the individual to the profile as the object on which truth is produced. The latter 'is a set of traits that does not pertain to a specific individual, but rather expresses relations between individuals, being more interpersonal than intrapersonal. Its main purpose is not to produce knowledge about an identifiable individual, but to use a set of personal details to act on those who are similar' (Bruno, 2013: 161).

As a result, the focus today is not on producing the truth about an individual's dangerousness, as in psychiatric assessment to detect abnormality in a subject, but rather the manifestation of truth is based on his profile, i.e. the statistical risk of him reoffending. Ruppert and Isin (2020) point out that this is a general logic of governance deployed through AI. The authors have coined the term 'sensory power' to refer to a new form of power that is no longer exercised over the individual, but over the clusters resulting from the data collected by the sensors that train the AI. The cluster, or risk profile, is thus a virtuality based on population factors that become real and individual in judicial decisions. Actuarial methods therefore produce a truth about the profile as if it were a truth about the individual. In this context, confession loses its privileged status as an act linking power-knowledge and subject, and gives way to a new kind of knotting based on profiling.

The perspective offered by Foucault's work on the link between power and truth provides an interesting approach to the role of AI in judicial processes. Indeed, it shows that

this modern technology has enabled the emergence of a new kind of link between power-knowledge and subject. However, it cannot provide a complete understanding of all the changes brought about by the use of AI in the judicial system. In fact, in addition to the above link, it reconfigures the number of players involved in truth production and the way they operate and interact. In order to analyse this second change more thoroughly, we must turn to the concept of exorganism proposed by Bernard Stiegler.

The Judicial Institution as an Exorganism

Foucault's work allows truth to be understood as an active and ongoing production process, yet it continues to focus on human beings as the sole source of truth. This fact, for example, changes radically in Bernard Stiegler's (2008, 2013, 2016, 2020) general organology. Stiegler's work is relevant in our case because it makes three particularly novel propositions. First, he argues that technology is nothing more than an exercise in the externalisation of human capacities and potentialities. This makes it a kind of new organ or exoorganism. Secondly, he argues that any exercise of knowledge production, such as a judicial verdict, cannot avoid what is mentioned in the first point, i.e. the use of technology must always be considered when attending to the aforementioned knowledge production. Finally, Stiegler's (2018, 2020) work distinguishes three levels of organs, which are inseparable because they are always associated: psychosomatic (the organic organ), technical (organological organs) and social (institutional organisms). These levels correspond to three types of externalisation of human potentialities. Thus, the human organism is always found within other social organisms, which in turn are associations of psychosomatic and technical organs. However, every social organism is an exorganism in that it requires organic organisms (human beings) associated with technical organs in order to reproduce and govern itself. Thus, it turns out that the judicial institution is a special exorganism that requires its association with organic organisms (such as judges, etc.) and technical organs (such as a polygraph) to produce truths. In the second decade of the 21st century, we note that a new technical organ, AI, has been added to this exorganism. It is true that Stiegler's work is considered a pessimistic proposal basically because of his insistence on the emergence of a new type of proletarianisation. This is defined as the standardisation of behaviour and the mechanisation of spirits through the externalisation of knowledge in technologies over which we have no control. However, this pessimism is nuanced by the notion of pharmakon.

Stiegler (2008, 2013, 2020) used the concept of *pharmakon* to conceptualise the distribution of agencies in exorganisms and the associations between psychosomatic and technical organs. Depending on how it is administered, *pharmakon* can be both cure and poison for the organic body. The author proposed a pharmacological interpretation by looking at associations with technology and showing in which circumstances this technology can be a cure for the psychosomatic organism and social organism and in which circumstances it acts as a poison. What is relevant in this notion is the insistence on the necessary interconnection between the levels mentioned above: the psychosomatic, technical and social levels are inseparable. Any small change in one of them, for example the introduction of AI in the production of judicial verdicts, implies a transformation in all the others and the creation of a new meaningful totality.

In order to put forward this idea, the author turned to Simondon's (2020) concept of transindividuation. For Simondon, the individual is always the result of a process of individuation in a system, combining pre-individual information (e.g. one's genetic code) with extra-individual information (e.g. one's relations with other individuals, objects, society or culture). Therefore, according to the author, to individuate is necessarily to individuate *with*, i.e. all individuation is transindividuation. This systemic relationship produces a metastable individual, since the individual is never finished and is forever in a process of transindividuation. From this perspective, Stiegler views the process of transindividuation as toxic when humans adapt to technology and lose agency in the governance of exorganisms. Conversely, when they embrace technology and retain agency in this governance, the association presents itself as a cure.

Thus, the combination of Stiegler's ideas with Foucault's proposal about truth allows us to analyse how judicial AI transforms the procedures of truth production, as well as its implications for human agency in the governance of these procedures. That is, in Stiegler's view, the use of AI in the judicial system involves, firstly, the realisation that non-human elements participate in the production of verdicts and new legal situations for human beings and, secondly and more importantly, that this use at the level of social externalisation changes the other two immediately. To explore these questions, we will turn to the three procedures of truth production described by Foucault.

Three Ways in Which AI Produces Judicial Truth

Analysing Sophocles' *Oedipus the King* as an example, Foucault (2014a, 2014b) distinguishes three forms in which subjects participate in the production of truth: as *operators*, in that truth flows through them by means of concrete operations (this is the case, for example, of priests, judges and scientists); as *spectators*, in that they bear witness to a reality that takes place before their eyes; and as *objects* of truth, in that the manifested truth is about them (Foucault, 2014a). Following Foucault's analysis, we intend to analyse the role of AI as a technical organ in the production of judicial truth through these three forms of subject participation.

AI as an Operator of Judicial Truth

As noted above, one of the functions of the subject in the alethurgy of truth manifestation described by Foucault is that of operator, i.e. that which performs a series of actions whereby truth flows through it. AI associates an organ in the role of operator in various phases of judicial alethurgy: choosing and assigning charges, handling simple cases and assisting judges when handing down prison sentences. Its action as an operator consists of processing data of different kinds using a variety of statistical machine learning and neural network models, each designed to perform optimally in specific situations (Park and Chai, 2021; Yang, 2021). Like the oracle in *Oedipus the King*, the truth emerging from patterns and correlations in reality flows through AI and is announced by it in the form of rankings and scores.

While judges and psychiatrists base their authority to formulate risk-related truths on the knowledge conferred to them by their rank and the use of certain tools (such as

psychological tests or written law), the authority of AI comes from the quality of the algorithmic models it applies. The quality of AI models is mainly measured by means of four metrics (He et al., 2018; Wang et al., 2020; Yan et al., 2019): precision, which measures the number of true positives with respect to the total number of positively classified cases; accuracy, which measures the total number of true positives and negatives with respect to the total number of classifications; recall, which measures the total number of true positives with respect to the sum of true positives and false negatives; and the F1 metric, which combines the precision and recall metrics. Here it should be noted that AI as an operator subject does not manifest the fairest truth but the most statistically probable one (Battelli, 2020).

AI is also the operator of a new kind of truth: a judicial meta-truth, i.e. a truth about how judicial truth manifests itself. This refers to AIs that are used to predict the outcome of court cases or the voting of individual judges. They are mainly used by law firms, Lex Machina being one good example, but models have also been developed by scholars for the same purpose. These tools predict the outcome of court cases based on extrajudicial factors such as the tone of voice in hearings (Henderson, 2014), the use of specific words in court records (Aletras et al., 2016; Lauderdale and Clark, 2014), the time and day of hearings (Chen and Eagel, 2017), the geographical and temporal proximity of cases (He et al., 2018), and public and lobbyist opinion (Sim et al., 2015). In this way, AI as an organ of knowledge produces a new truth in which the judicial organism appears as an object of knowledge, reshaping the agency of human subjects insofar as it detracts from their authority. However, it also presents as a useful organ for the detection and correction of biases within the legal profession (Battelli, 2020). In fact, even though the judicial process itself becomes an object of truth, it can still be understood in terms of risk: the statistical calculation of the probability that an event, in this case a conviction, will occur.

Finally, it is worth noting that AI as an organ associates agents such as judges or psychiatrists, but also programmers and designers, who must make decisions on how to translate certain categories into algorithmic processes. Such categories include fairness and equal treatment (Berk et al., 2021; Li et al., 2022; Zhang et al., 2019), the establishment of causality (Zennaro and Ivanovska, 2018) and the transparency of algorithmic processes (Pei et al., 2019). In turn, there is a restructuring of the procedures, agencies and entities through which truth is manifested in the judicial organism. Thus, with the introduction of AI, the operator of judicial truth is an exorganism resulting from the association of human agents and algorithmic processes.

AI as Spectator to Dividual Truth

Judicial AIs primarily use natural language processing (NLP), which is responsible for transforming all judicial texts (court records, sentences, verdicts, laws, precedents, etc.) into mathematical patterns in terms of frequency, length and location of topics (word sets) and correlating them, for example, with the charges brought against the accused and the court decision. NLP takes the text from its fragments, its dividuality, which allows AI to witness truths other than those of the text taken as an individual unit. As Henderson (2014) argued, NLP as a tool implies a ‘distant reading’ of the text in the sense that it

does not look for meanings but for statistical patterns and differs from the judge's reading, which seeks to create meaning and a narrative between law, precedents and facts.

For example, Li et al. (2020) presented an AI that assists in the process of choosing the charges to be brought by the prosecution based on the facts of the crime. It is a neural network with three main modules: a fact encoder, a crime interpretation encoder and a charge predictor. The encoder modules are responsible for transforming court texts into word vectors with which to detect key topics that correlate statistically with concrete facts and concrete crimes, respectively. Combining the results of both encoders with the charges that have been handed down in each case, the last module generates charge predictions. The process is more complex than described here, but it serves to illustrate how this AI does not take texts as individual and singular entities, but interprets them at the dividual level: from their fragments (word topics) and their relation to other text fragments (correlations of topics between facts, crime descriptions and charges).

Therefore, if we look at the role of AI in the judicial organism as a spectator to truth – a truth that does not flow through it, but which it observes externally – we see that it witnesses a truth that goes beyond the sentience of the human psychosomatic organ of knowledge (Pasquinelli and Joler, 2020). We are referring to the ability to find patterns on the plane of what Deleuze called the dividual (Deleuze and Guattari, 2015). That is, entities taken not as individual units but as fragments of these individualities. Using sensors and algorithmic processes, AI witnesses these infra-individual fragments (e.g. gender, race, place of residence, topics in a text, tone of voice and the time at which a hearing occurs), with which it detects patterns on a supra-individual plane (Raunig, 2016; Rodríguez, 2019; Rouvroy and Berns, 2013). This process is called profiling or profile generation, as explained above.

It is important to note that we are not just talking about profiles created on the basis of human individuals (as with Netflix or Spotify profiles). Rather, we can extend the definition to any type of entity. Thus, we can speak of profiles of an individual's risk of recidivism, but also profiles of sentences, fines, tones of voice, crimes or hearings. What counts most in this definition is the emphasis on the fact that profiles are generated by witnessing the relationships between pre-individual fragments or traits, by picking up on their links at the dividual level (Bruno and Rodríguez, 2021; Lury and Day, 2019), and that their main purpose is to calculate the probability of a factor manifesting itself within a set of variables. By shifting the observed truth to the dividual level, we also see that the subjects' actions and statements are relegated to a secondary level: they are collected as fragments in the form of data, but the truth emerges from their relationality and not from the concrete act of the individual (Rouvroy and Berns, 2013).

Other similar examples of this observed truth at the textual dividual level include Park and Chai's (2021) verdict prediction AI for invasion of privacy lawsuits, Wang et al.'s (2020) indictment predictor, and Lauderdale and Clark's (2014) US Supreme Court politician vote predictor, to name a few. In judicial truth production, this distant reading displaces or transforms Western regimes of truth, which have been primarily concerned with writing and meaning (Rouvroy and Stiegler, 2016; Stiegler, 2016). It is not our intention to claim that this displacement in the Western truth regime is completely new or that it has been brought about by AI. For example, such displacement had already occurred in the field of microbiology with disease prediction according to genetic

patterns (Rodríguez, 2019). Our point, rather, is that this displacement is ongoing and, with the emergence of AI as an organ, takes on another dimension.

Therefore, in the production of judicial truth, we find different types of organs, both organic and organological, which are associated into a truth-witnessing exorganism: eyewitnesses, camera recordings, depositions, police reports, recorded telephone conversations, and so on. AIs appear as an additional organ that is capable of witnessing, but the truth they witness differs from the abovementioned in that it is situated within the relations between fragments of things or facts and not in the things or facts themselves as units.

AI as Object of Truth

If we turn our attention to the third function that Foucault assigns to the subject in the process of manifesting truth, that of object, we find that AI as an organ also plays this role in judicial truth production. This applies to the procedures to which the AI is subjected in order to ascertain the truthfulness of its statements. This is achieved through previously mentioned metrics such as precision, accuracy, recall and F1 (a measure of the harmonic mean of precision and recall), but also through other more complex metrics, such as those that make it possible to determine the model's level of fairness. These metrics are obtained from algorithmic operations that AI performs on the data before and during processing or on the results thereof (Berk et al., 2021).

Precision, accuracy, recall and F1 are commonly used to assess the quality of AI results, i.e. how reliable the statistical predictions (the truth) it delivers are. For example, Park and Chai (2021) compared the accuracy of five algorithmic classification models (LDA, Red Neuronal, CART, SVM and Random Forest), finding that CART is the best-performing model in the application of network text analysis (NTA), the statistical procedure used to process court texts and predict sentences in invasion-of-privacy lawsuits. Yan et al. (2019) also used these four metrics to show that the TextCNN model performs best at predicting the main articles of law related to a specific case.

However, several authors agree that these metrics are not enough to assess the quality of judicial AIs because fairness is also a key issue to consider in this field. For example, when calculating an offender's risk of recidivism, the AI must not only be precise and accurate, but also treat all groups equally and non-discriminatorily. This in turn begs the question of how fairness and equity are to be defined and algorithmically translated. Yanjun Li et al. (2022) posited that a model M can be considered fair with respect to an attribute A when two requirements are met: (1) the absence of attribute A in a dataset D does not affect the performance of model M; and (2) the value of attribute A in a dataset D does not affect the performance of model M, even when this value is counterfactual. The authors found that machine learning models are fairer than neural networks, but the latter have a higher degree of accuracy. However, they noted that accuracy is also related to making fair decisions, so a balance must be struck between the two variables.

The same conclusion was drawn by Berk et al. (2021), who defined fairness and equity in terms of the model performing the same percentage of misclassifications across all groups. This is achieved by programming the AI to be less accurate with the most strongly correlated groups, for which they presented five procedures. The authors noted

that each procedure comes with its own specific problems. For example, considering false positives and false negatives equally is problematic in some situations, such as pre-trial detention cases, as the procedure would value letting a repeat offender go free the same as imprisoning an innocent person, and this is questionable. In addition to the limitations of each procedure, some of them are contradictory and exclusionary, as reducing the severity of the AI with some groups that are more highly correlated implies a loss of accuracy. Therefore, the authors concluded that it is not possible to achieve an AI that is both fair and accurate.

Other similar cases are the study by Zhang et al. (2019), who used three metrics to assess the robustness of models when generating mutations in their databases, and Pei et al.'s (2019) algorithm for testing neural networks to ensure that they take into account the so-called black swans or corner cases. The point of presenting these examples is not to analyse the concrete algorithmic processes used in judicial AIs but to show that AI is an object organ about which truth is produced in terms of how reliable or credible its statements are.

AI itself also participates in this search for truth, transforming itself through recursive self-operations. Recursivity, as Hui (2019) noted, is not mere mechanical repetition, but a looping movement open to transformation. Thus, an AI can be programmed to recursively apply operations onto itself by varying algorithmic models until it finds the combination that yields the best metrics. In this sense, judicial AI can itself be seen as an exorganism, one that runs the risk of governing itself through automatisms, thus depriving humans of agency. AI understood as an exorganism would be associated with the judicial exorganism, increasing the complexity of maintaining human agency in the production of truth.

Discussion and Conclusions

All of the above gives us three main points to consider. The first point is that the judicial institution is an exorganism composed of various human-machine associations and that judicial truth is contingent on this reality. In this regard, we have shown that judicial truth comprises various phases or procedures in which AI intervenes as an organ in the role of operator, spectator and object. As an operator, judicial AIs carry out a series of algorithmic procedures whereby truth flows through them. They state truths based on the objectivity of the statistical models they implement and not on a qualification or acquired knowledge, as is the case with a judge or psychiatrist. The operator of judicial truth can be regarded as an organism in which judges are associated with a new technical organ that assists them by providing a type of knowledge that differs from their own as human-experts. We have also pointed out that some AIs are operators of a truth about judicial truth, stating truth about how judges, in turn, produce truths. In this case human agency is diminished and reduced to the role of mere bias corrector.

By analysing judicial AI as a spectator, we have shown that it tells a truth that is outside of itself and that has nothing to do with entities taken as individual units, as is the case with human beings. Rather, it operates on the basis of entities defined by their infra-individual fragments in order to observe links on a supra-individual level. This refers to the profiling of both human and non-human entities, shifting truth production from the

individual to the profile. In this respect, we have pointed to NLP as a prime tool for AIs, as it enables them to read judicial texts at the dividual level. This, in turn, allows them to witness truths through statistical correlations between word topics rather than from word meanings.

In the case of judicial AI as an object about which truth is manifested, we have seen that a series of mathematical operations are performed on it, producing metrics that speak to the AI's truth-telling, i.e. about the truthfulness and certainty attributable to the algorithmic model in question. The application of these metrics requires an association with human psychosomatic organs. This is because they have to be translated into algorithmic operations, which involves creating mathematical definitions of the concepts and selecting some procedures over others. However, we have also shown that AI can autonomously optimise its own models by recursively applying these metrics and procedures. It can therefore be seen as an exorganism itself, over which human governance must be ensured.

Secondly, it follows from the above that the conceptualisation of AI as another organ in the judicial exorganism reshapes the distribution of agencies in the production of truth. This challenges the position of humans as the main source of truth and displaces the Western regime of truth manifestation that is mainly based on texts and meaning (Stiegler, 2016). However, this displacement should not be construed as replacing humans with technology, but rather as transforming the organs of knowledge. This has great potential, provided that we ensure that human agency is not totally nullified. As Stiegler's (2013) pharmacology approach contends, we can only have that security if human subjects maintain the capacity to transindividuate with judicial AIs. That is, they must appropriate and embrace them as part of the judicial exorganism, as organs that fulfil some function that complements human agency: speeding up processes, detecting biases or finding correlations that are undetectable to the human eye, for example. Authors such as Rizer and Watney (2018) have pointed out that judges must urgently be trained to interact with and interpret AIs in order to avoid blindly following their recommendations.

Finally, we should highlight the potential toxicity of AI as an organ of judicial truth production for the population as a whole. With the inclusion of judicial AI, risk profiles not only include sensitive categories such as gender and race; depending on the data fed into the AI, they could also contain traits that have no inherent relevance to the judicial sphere, such as consumer habits. Moreover, as Harcourt (2007) points out, the use of statistical methods leads to a self-perpetuating targeting of certain specific profiles, as it tends to increase the harshness of sentences for these groups and overlook the risks of others. The inclusion of AI exacerbates this problem, because the more conviction data the AI accumulates about particular profiles, the harsher its outcomes will be towards those profiles. However, as we have noted, actuarial methods are undergoing an even more profound change: a shift from the individual to the profile in the production of truth about the risk of committing crime. By transforming virtual events into real ones, this shift entails a loss of human agency, in that the individual loses his ability to generate truth about himself. In this way, he is unable to resist his subjectification as a criminal, as traits that statistically belong to his profile are taken as if they were his own (Rouvroy and Berns, 2013). Couldry and Mejías (2019) note that this implies a violation of the minimal integrity of the self, understood as the minimum conditions that allow an

individual to generate a horizon of imagination and action. Thus, the totalising truth of the profile undermines the autonomy and freedom of the individual, that is, his agency to open up a space of possibilities around himself.

For all these reasons, we can conclude that AI actively participates in the judicial exorganism, manifesting truths at the dividual level and expanding the objects on which truth is produced. Although judicial AI speeds up sentencing and facilitates decision-making, it is no less true that it displaces judges, experts and individuals in the production of truth about their actions, insofar as it installs the profile as an essential part or central loop of the knot between truth-power and subject. Following Stiegler's words, we could say that AI unfolds like a pharmakon in our most immediate reality: it facilitates some courses of action and activities, yet opens up others that may lead to situations that we now consider unjust or inappropriate.

We would like to end this text by commenting that the dangers pointed out are framed within the logic of the liberal subject that places its centre of interest in guaranteeing autonomy, equity and equality. However, there are perspectives that, without undermining the importance of these issues, propose thinking beyond the emancipation of the liberal subject (Petersmann and Van Den Meerssche, 2024). Specifically, they question the possibility of not using the profile in order to associate the most probable event with the individual, but to open up possible futures that contribute to collective emancipation. Although there is no answer to this question, we believe it is a line of thought that should be addressed in future research.

Funding

This paper was written as part of the research project 'Future Scenarios of Local Development Policies in the Post-Pandemic Era: From Biopolitics to Psychopolitics', which is part of a broader project entitled 'Power and Control Relationships in Citizen Science'. The research was carried out through the Barcelona Science and Technology Studies Group (STS-b) under the doctoral programme in Person and Society in the Contemporary World, Department of Social Psychology, Universitat Autònoma de Barcelona (Spain).

ORCID iD

Joan Rovira Martorell https://orcid.org/0000-0001-5638-1808

Notes

1. PredPol is an AI from the company PredPol that uses machine learning models to process three types of data: type of crime, location, and date and time.
2. Víctor is an AI used by the Brazilian Supreme Federal Court to determine which petitions meet the minimum requirements to merit consideration.
3. PretorIA is an AI used by the Constitutional Court of Colombia to classify applications for state protection in the field of health according to criteria set by the court.
4. HART is an AI developed by the University of Cambridge to predict a criminal's risk of recidivism in the next two years.
5. COMPAS is an AI developed by Equivant to create psychological profiles of defendants based on personal data (some of which is considered sensitive) in order to generate risk scales in three areas: risk of pretrial release, general risk of recidivism and risk of violent recidivism.

6. Prometea is an AI developed by the Public Prosecutor's Office of the Autonomous City of Buenos Aires to autonomously resolve public matters such as the enforcement of fines or the allocation of licences.
7. Lex Machina is an AI developed by Lexis Nexis to advise law firms on their client selection strategy by predicting the outcome of litigation.

References

Aletras, Nikolaos, Tsarapatsanis, Dimitrios, Preoțiuc-Pietro, Daniel, et al. (2016) Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2: e93.

Anderson, Chris (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 23 June. Available at: https://www.wired.com/2008/06/pb-theory/ (accessed 25 July 2024).

Battelli, Ettore (2020) La decisión robótica: Algoritmos, interpretación y justicia predictiva. *Revista de Derecho Privado* 40: 45–86.

Berk, Richard, Heidari, Hoda, Jabbari, Shahin, et al. (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50(1): 3–44.

Bex, Floris and Prakken, Henry (2021) Can predictive justice improve the predictability and consistency of judicial decision-making? In: Schweighofer, E. (ed.) *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press.

Bruno, Fernanda (2013) *Máquinas de ver, modos de ser: vigilância, tecnologia e subjetividade*. Porto Alegre: Sulina.

Bruno, Fernanda and Rodríguez, Pablo Manolo (2021) The dividual: Digital practices and biotechnologies. *Theory, Culture & Society* 39(3): 1–24.

Casabona, Carlos María Romeo (2018) Riesgo, procedimientos actuariales basados en inteligencia artificial y medidas de seguridad. *Revista de Derecho, Empresa y Sociedad* 13: 39–55.

Chen, Daniel L. and Eagel, Jess (2017) Can machine learning help predict the outcome of asylum adjudications? In: *Proceedings of ACM ICAIL Conference on AI and the Law*, London, June 2017. New York: ACM Publishing, pp. 1–12.

Couldry, Nick and Mejías, Ulises A. (2019) *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.

Deleuze, Gilles and Guattari, Felix (2015) *A Thousand Plateaux*. Edinburgh: Edinburgh University Press.

Ewald, François (1991) Insurance and risk. In: Burchell, Graham, Gordon, Colin, and Miller, Peter (eds) *The Foucault Effect*. Chicago: University of Chicago Press, pp. 197–210.

Foucault, Michel (1979) *Microfísica del poder*. Madrid: Las Ediciones de La Piqueta.

Foucault, Michel (1981) The order of discourse. In: Young, Robert (ed.) *Untying the Text: A Post-Structuralist Reader*. Boston: Routledge & Kegan Paul Ltd, pp. 48–78.

Foucault, Michel (1996) *La verdad y las formas jurídicas*. Barcelona: Gedisa.

Foucault, Michel (2003) *Abnormal: Lectures at the Collège de France, 1974–1975*. London: Verso.

Foucault, Michel (2008) *Tecnologías del yo y otros textos afines*. Buenos Aires: Paidós.

Foucault, Michel (2009) *Security, Territory and Population: Lectures at the Collège de France, 1977–1978*. London: Picador.

Foucault, Michel (2010) *The Birth of Biopolitics: Lectures at the Collège de France, 1978–1979*. London: Picador.

Foucault, Michel (2014a) *On the Government of the Living: Lectures at the Collège de France, 1979–1980*. London: Palgrave Macmillan.

Foucault, Michel (2014b) *Wrong-Doing, Truth-Telling: The Function of Avowal in Justice*. Chicago: University of Chicago Press.

Harcourt, Bernard E. (2007) *Against Prediction: Profiling, Policing and Punishing in the Actuarial Age*. Chicago: University of Chicago Press.

He, Tie-Ke, Lian, Hao, Qin, Ze-Min, et al. (2018) PTM: A topic model for the inferring of the penalty. *Journal of Computer Science and Technology* 33(4): 756–767.

Henderson, William D. (2014) From big law to lean law. *International Review of Law and Economics* 38: 5–16.

Hui, Yuk (2019) *Recursivity and Contingency*. Lanham: Rowman & Littlefield.

Jin, Yaohui and He, Hao (2020) An artificial-intelligence-based semantic assist framework for judicial trials. *Asian Journal of Law and Society* 7(3): 531–540.

Lauderdale, Benjamin E. and Clark, Tom S. (2014) Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science* 58(3): 754–771.

Levmore, Saul and Fagan, Frank (2019) The end of bargaining in the digital age. *Cornell Law Review* 103(6): 1469–1526

Li, Xinchuan, Kang, Xiaojun, Wang, Chenwei, et al. (2020) A neural-network-based model of charge prediction via the judicial interpretation of crimes. *IEEE Access* 8: 101569–101579.

Li, Yanjun, Huang, Huan, Geng, Qiang, et al. (2022) Fairness measures of machine learning models in judicial penalty prediction. *Journal of Internet Technology* 23(5): 1109–1116.

Lury, Celia and Day, Sophie (2019) Algorithmic personalization as a mode of individuation. *Theory, Culture & Society* 36(2): 17–37.

Park, Minjung and Chai, Sangmi (2021) AI model for predicting legal judgments to improve accuracy and explainability of online privacy invasion cases. *Applied Sciences* 11(23): 11080.

Pasquinelli, Matteo and Joler, Vladan (2020) The nooscope manifested: AI as instrument of knowledge extractivism. *AI & Society* 36: 1263–1280.

Pei, Kexin, Cao, Yinzhi, Yang, Junfeng, et al. (2019) DeepXplore: Automated whitebox testing of deep learning systems. *Communications of the ACM* 62(11): 137–145.

Petersmann, Marie, and Van Den Meerssche, Dimitri (2024) On phantom publics, clusters, and collectives: Be(com)ing subject in algorithmic times. *AI & Society* 39(1): 107–124.

Raunig, Gerald (2016) *Dividuum: Machinic Capitalism and Molecular Revolution*. South Pasadena: Semiotext(e).

Rizer, Arthur and Watney, Caleb (2018) *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable and Just* (SSRN Scholarly Paper no. 3129576).

Rodríguez, Pablo Manolo (2019) *Las palabras en las cosas. Saber, poder y subjetivación entre algoritmos y moléculas*. Ciudad Autónoma de Buenos Aires: Editorial Cactus.

Rouvroy, Antoinette and Berns, Thomas (2013) Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation through relationships? *Réseaux* 177: 163–196.

Rouvroy, Antoinette and Stiegler, Bernard (2016) The digital regime of truth: From the algorithmic governmentality to a new rule of law. *La Deleuziana* 3: 6–29.

Ruppert, Evelyn and Isin, Engin (2020) The birth of sensory power: How a pandemic made it visible. *Big Data & Society* 7(2): 1–15.

Sadin, Eric (2020) *La inteligencia artificial o el desafio del siglo. Anatomía de un antihumanismo radical*. Ciudad Autónoma de Buenos Aires: Caja Negra Editora.

Sim, Yanchuan, Routledge, Bryan, and Smith, Noah (2015) The utility of text: The case of Amicus briefs and the Supreme Court In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, 25–30 January 2015. Palo Alto: AAAI Press, pp. 2311–2317.

Simondon, Gilbert (2020) *Individuation in Light of Notions of Form and Information*. Minneapolis: University of Minnesota Press.

Stiegler, Bernard (2008) *Technics and Time, 2: Disorientation*. Stanford: Stanford University Press.
Stiegler, Bernard (2013) *What Makes Life Worth Living: On Pharmacology*. Cambridge: Polity Press.
Stiegler, Bernard (2016) *Automatic Society, Vol. 1: The Future of Work*. Cambridge: Polity Press.
Stiegler, Bernard (2018) *The Neganthropocene*. London: Open Humanities Press.
Stiegler, Bernard (2020) *Nanjing Lectures (2016–2019)*. London: Open Humanities Press.
Ulenaers, Jasper (2020) The impact of artificial intelligence on the right to a fair trial: Towards a robot judge? *Asian Journal of Law and Economics* 11(2): 1–38.
Wang, Yongjun, Gao, Jing and Chen, Junjie (2020) Deep learning algorithm for judicial judgment prediction based on BERT In: *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, 14–16 October 2020, Patna, India. Piscataway: Institute of Electrical and Electronic Engineers, pp. 1–6.
Yan, Ge, Li, Yu, and Shen, Siyuan, et al. (2019) Law article prediction based on deep learning In: *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 22–26 July 2019, Sofia, Bulgaria. Piscataway: Institute of Electrical and Electronic Engineers, pp. 281–284.
Yang, Juan (2021) Analysis and application of judicial data based on machine learning. *Journal of Physics: Conference Series* 1846(2021): 012027.
Zennaro, Fabio Massimo and Ivanovska, Magdalena (2018) *Pooling of Causal Models under Counterfactual Fairness via Causal Judgement Aggregation*. DOI: 10.48550/arXiv.1805.09866.
Zhang, Shu, Yan, Ge, Li, Yu, et al. (2019) Evaluation of judicial imprisonment term prediction model based on text mutation. In: *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 22–26 July 2019, Sofia, Bulgaria. Piscataway: Institute of Electrical and Electronic Engineers, pp. 62–65.

Joan Rovira Martorell is a PhD student under the doctoral programme in 'Person and Society in the Contemporary World', Department of Social Psychology, Universitat Autònoma de Barcelona. His main line of research is the link between power, politics and artificial intelligence.

Ana Gálvez is Senior Lecturer of Social Psychology at the Universitat Oberta de Catalunya. She is part of the STS-b group (Barcelona Science and Techology Studies Group) and her main lines of research have to do with gender and new work modalities.

Francisco Tirado is Senior Lecturer of Social Psychology at the Universitat Autònoma de Barcelona. He is part of the STS-b group (Barcelona Science and Techology Studies Group). His research interests cover three broad areas: a) social studies of science and technology; b) power relations and political action in new sociotechnical contexts; c) the transformation of biopower in psychopower.