DOI: 10.14515/monitoring.2024.5.2580

A. P. Kazun

CAN ARTIFICIAL INTELLIGENCE PREDICT JUDICIAL DECISIONS? A SYSTEMATIC REVIEW OF INTERNATIONAL RESEARCH

For citation:
Kazun A. P. (2024) Can Artificial Intelligence Predict Judicial Decisions? A Systematic Review of International Research. *Monitoring of Public Opinion: Economic and Social Changes.* No. 5. P. 100–122. https://doi.org/10.14515/monitoring.2024.5.2580. (In Russ.)
Правильная ссылка на статью:
Казун А. П. Может ли искусственный интеллект прогнозировать решения суда? Систематический обзор международных исследований // Мониторинг общественного мнения: экономические и социальные перемены. 2024. № 5. С. 100—122. https://doi.org/10.14515/monitoring.2024.5.2580.

Получено: 05.03.2024. Принято к публикации: 15.08.2024.

DOI: 10.14515/monitoring.2024.5.2580

CAN ARTIFICIAL INTELLIGENCE PREDICT JUDICIAL DECISIONS? A SYSTEMATIC REVIEW OF INTERNATIONAL RESEARCH

Anton P. KAZUN[1] *— Cand. Sci. (Soc.), Director at the Institute for Industrial and Market Studies; Assistant Professor at the Department of Applied Economics*
E-MAIL: akazun@hse.ru
https://orcid.org/0000-0002-0091-5388

[1] HSE University, Moscow, Russia

Abstract. Advancements in artificial intelligence technologies and the emergence of open databases containing judicial decisions have led to rapid improvements in algorithms capable of classifying legal documents and forecasting decisions made by judges. This article examines a body of international research dedicated to how accurately AI can predict judges' decisions and whether it could potentially replace human judges in the future. The answer to this question is formed by analyzing two key aspects: the capability and accuracy of predicting judicial decisions and the various constraints associated with using AI.

Analysis of international experience shows that the accuracy of predictions has increased in recent years; however, the quality of the models depends greatly on the specificity of the tasks and the available data. Most studies analyze decisions from higher courts worldwide, significantly reducing their practical potential for dealing with mass categories of cases. Moreover, concerns have arisen regarding the use of models that operate on a "black box" principle, as their decisions are difficult to interpret. Despite the rapid development of AI technologies, the complete replacement of judges is unlikely because of the range of methodological limitations, including insufficient quality and volume of data, issues with interpretability,

МОЖЕТ ЛИ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ПРОГНОЗИРОВАТЬ РЕШЕНИЯ СУДА? СИСТЕМАТИЧЕСКИЙ ОБЗОР МЕЖДУНАРОДНЫХ ИССЛЕДОВАНИЙ

КАЗУН Антон Павлович — кандидат социологических наук, директор Института анализа предприятий и рынков, доцент департамента прикладной экономики, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: akazun@hse.ru
https://orcid.org/0000-0002-0091-5388

Аннотация. Развитие технологий искусственного интеллекта и появление открытых баз данных судебных решений привели к стремительному совершенствованию алгоритмов, позволяющих классифицировать юридические документы и прогнозировать принимаемые судьями решения. В статье мы анализируем корпус международных исследований, посвященных вопросу о том, насколько точно ИИ может предсказывать решения судей и, как следствие, сможет ли он в перспективе заменить судью-человека. Ответ на этот вопрос складывается из анализа двух ключевых аспектов — возможности и точности прогнозирования судебных решений, а также различных ограничений, связанных с применением ИИ.

Анализ международного опыта показывает, что в последние годы точность прогнозов выросла, однако качество моделей сильно зависит от специфики задач и доступных данных. Большинство исследований анализируют решения судов высшего уровня различных стран мира, что сильно снижает их прикладной потенциал в части работы с массовыми категориями дел. Кроме того, опасения вызывает использование моделей, действующих по принципу «черного ящика», поскольку их решения трудно интерпретировать. Несмотря на стремительное

DOI: 10.14515/monitoring.2024.5.2580

challenges in understanding legal and cultural context, and limitations in transferring models to other legal systems. However, AI technologies can be used to reduce the costs associated with case material handling.

развитие ИИ-технологий, полная замена судей вряд ли возможна в ближайшее время ввиду целого ряда методологических ограничений, включая недостаточное качество и объем данных, проблему интерпретируемости, сложность понимания юридического и культурного контекста, ограничения переноса на другие правовые системы. Однако ИИ-технологии возможно использовать для сокращения издержек по работе с материалами дела.

Keywords: artificial intelligence, prediction of judicial decisions, machine learning, deep learning, legal document classification, algorithmic accuracy

Ключевые слова: искусственный интеллект, предсказание судебных решений, машинное обучение, глубокое обучение, классификация юридических документов, точность алгоритмов

Acknowledgments. The research was funded by the Russian Science Foundation grant No. 23-78-10073 "Development and approbation of the algorithm for automated analysis of the court decisions texts for socio-legal studies (based on cases of violent crimes)" (see more: https://rscf.ru/project/23-78-10073/).

Благодарность. Исследование выполнено за счет гранта Российского научного фонда № 23-78-10073 «Разработка и апробация методики автоматизированного анализа текстов приговоров российских судов для социально-правовых исследований (на примере насильственных преступлений)» (см. подробнее: https://rscf.ru/project/23-78-10073/).

Introduction

Since the 1950s, researchers have been trying to predict judges' decisions using quantitative methods [Kort, 1957], but until recently, technology has only been able to achieve very limited results in this area [Ashley, Brüninghaus, 2009]. In recent years, with the development of artificial intelligence (hereafter AI) technologies and the increasing availability of open court data, the quality of models predicting judges' decisions has increased significantly [Medvedeva, Wieling, and Vols, 2022]. This has led to a widening debate on whether AI can replace the professional judge or at least simplify the work of judges in the foreseeable future [Shi, Sourdin, Li, 2021; Sourdin, Cornes, 2018; Taylor, 2023; Xu, 2022]. From 2017 to 2022, the number of law-related papers presented at specialized computer science conferences increased dramatically [Cui, Shen, Wen, 2023]. In this study, based on a systematic analysis of international experience, we attempt to describe the opportunities and risks brought about by the emergence of AI technology in the field of judicial proceedings, and answer the question of whether it is possible to replace the judge with AI. It should be noted that in this article we will talk

DOI: 10.14515/monitoring.2024.5.2580

exclusively about so-called «narrow AI» that can solve specific tasks (in our case, analyzing legal texts and predicting outcomes), leaving aside the issue of creating «strong» or general artificial intelligence capable of reproducing the entire human thought process.

For the legal field, the discussion of AI implementation is perhaps no less relevant than that for medicine, as in both cases, the lives and health of people are often at stake. AI is capable of analyzing legal texts in a volume and at a speed that is beyond human capabilities. However, the «costs of error» of introducing AI into the judicial sphere, or more broadly, into law enforcement might be high. For example, in the 2001 movie *Minority Report* the main character, who worked for a corporation that predicted serious crimes before they occurred, became a victim of the algorithm himself when he was accused of a crime he did not commit. Although we are a long way from such a future, with the development of AI, the ability to predict crimes in a statistical sense has become quite realistic [Gerber, 2014]. In addition to the identified ethical issues, the widespread adoption of AI has raised other questions. For example, this practice poses the problem of protecting confidential information and raises the issue of the privacy of ordinary citizens' data. However, on the other hand, it is possible to significantly reduce the costs of judicial decision-making and possibly improve the quality and impartiality of justice.

In the Russian context, a discussion of the implications of the introduction of AI in law enforcement is also relevant, as the development of this technology is a national priority. In January 2024, Vladimir Putin instructed the Supreme Court, the Prosecutor General's Office, the Investigative Committee, the Interior Ministry, and the Ministry of Justice of the Russian Federation to consider the possibility of using AI in the investigation of crimes [1]. In the Address to the Federal Assembly of February 29, 2024, instruction was given to increase the capacity of domestic supercomputers tenfold by 2030 [2]. Although the first initiatives in the field of law enforcement concerned the fight against crime using video cameras and facial recognition technology, there is no doubt that AI will be used in judicial practice as well. In particular, experiments to create «smart courts» are already underway in China [Rusakova, 2021; Shi, Sourdin, Li, 2021]. Some Russian judges were skeptical about the prospect of their replacement by AI [3], however, the launch of the *Justice Online* system [4] was announced for 2024, which would make the first step towards the automation of typical components of the judicial process. In addition, in the summer of 2024, Judge Oleg Zatelepin of the Supreme Court of the Russian Federation suggested that AI could be introduced into the Russian justice system to help judges make de-

[1] Putin instructed to improve the use of AI to investigate crimes // RBC. 2024. January 17. URL: https://www.rbc.ru/rbcfreenews/65a7a3359a794761a6913056(Accessed on 05.03.2024).

[2] Putin instructed to increase the capacity of supercomputers ten times // RBC. 2024. February 29. URL: https://www.rbc.ru/rbcfreenews/65e07a3a9a79472386b14ad3(Accessed 05.07.2024).

[3] The Council of Judges rejected the idea of replacing a judge with an AI // Pravo.ru. 2023. October 26. URL: https://pravo.ru/news/249529/(Accessed 05.07.2024).

[4] Kulikov V. Courts are planning to connect artificial intelligence to the drafting of decisions // Rossiyskaya Gazeta. 2023. May 25. URL: https://rg.ru/2023/05/25/robot-pomozhet-rassudit.html(Accessed 05.07.2024).

DOI: 10.14515/monitoring.2024.5.2580

cisions on punishment and to reduce the probability of errors [5], but only in the role of an advisor or assistant to the judge.

Since most Russian judges are overloaded with work [6], the introduction of AI technologies to analyze case files could at least help free up time for meaningful familiarization with their key circumstances. For lawyers and their clients, the emergence of such technologies would greatly simplify the preparation for court proceedings, as well as provide an opportunity to assess the chances of success [Jacob de Menezes-Neto, Clementino, 2022].

This paper is structured as follows. First, we describe an algorithm for searching and working with relevant scientific literature on the issue of the AI-assisted prediction of judicial decisions. We then sequentially answer several key questions about the use of AI in categorizing legal texts and predicting judicial decisions. This part is followed by the discussion concering the limitations of the applicability of this technology, addressing key ethical issues associated with the implementation of AI, and providing practical recommendations.

Source selection and research questions

The selection of sources for this study was performed through Research Rabbit [7], an AI application that allows building citation networks between scientific sources. The advantages of this approach compared to manual literature search using keywords in the Scopus and WoS scientific citation databases are that the algorithm highlights the core of scientific discussion and is less dependent on the choice of keywords. As a starting point for building the network of sources, we used a sample from a systematic review [Medvedeva, Wieling, Vols, 2022], which selected relevant studies from 2015 to 2021 on the issue of predicting judgmental decisions. The use of Research Rabbit allowed us to bring this sample up to date as well as to find texts on related topics, including discussions of ethics. The accuracy of the selection of studies was further checked using a second resource, Inciteful [8]. It works in a similar manner and allows us to identify texts that are similar in content to the available sample. The approach we used, like the keyword search, does not exclude the possibility of missing a particular study, but the latter certainly does not make a significant contribution to the academic debate because of its lack of strong links to the main body of literature.

The final network included 107 sources from 2004 to 2023 on the classification of judicial texts, but most of the works were published since 2018, as the debate on the use of AI has intensified in recent years. In general, the obtained set of texts covers all of the most relevant sources dedicated to judicial decision prediction. To quantitatively analyze the quality dynamics of the prediction models, 34 sources

[5] Burnov V. VS: AI can help reduce the number of errors in sentencing // RAPSI. 2024. July 1. URL: https://www.rapsinews.ru/judicial_mm/20240701/310059533.html(Accessed 05.07.2024).

[6] The burden on judges will be fixed at the legislative level // Advokatelskaya Gazeta. 2023. December 06. URL: https://www.advgazeta.ru/novosti/nagruzku-na-sudey-zakrepyat-na-zakonodatelnom-urovne/(Accessed 05.07.2024).

[7] URL: https://www.researchrabbit.ai/(Accessed 05.07.2024).

[8] URL: https://inciteful.xyz/(Accessed 05.07.2024).

DOI: 10.14515/monitoring.2024.5.2580

were selected for the period from 2015 to 2023, comparable in terms of problem statements and accuracy of assessment metrics.

The analysis of foreign experience has several limitations. First, we do not analyze methodological issues in detail, which are the subject of a large body of literature in the field of computer science. In this study, we do not aim to describe existing algorithms for classifying legal texts or to reveal the ways in which they are applied. This limitation, among others, allowed us to stop building the network in Research Rabbit when new related texts ceased to be concerned with predicting judicial decisions and began to address related topics in mathematics and computer science. Second, we do not analyze papers dealing with predictions in law enforcement and crime, focusing only on the court case stage. This allows us to maintain a clear research focus and not to stray into other complex issues, such as the use of AI as a tool to find criminals, a topic of high research interest, but a subject for a separate analysis. Third, we do not examine in detail papers modeling judicial decisions based on linear regression models, decision trees, and other classical quantitative methods, as these studies have given way in recent years to more advanced methods based on modern AI technologies. Simultaneously, the analysis includes articles devoted to ethical aspects of AI application in law enforcement, which were not directly related to decision prediction.

The analysis of literature sources in Russian has shown that, despite the presence of quite a large number of studies [Zakhartsev, Salnikov, 2018; Kovalenko et al., 2020; Kravchuk, 2021; Stepanov, Basaganov, 2022], they offer mainly the statement of topical issues and analyze the legal aspects of the problem, but do not systematize the international experience of using AI technologies to predict judicial decisions. Thus, this article is the first systematic analysis of a corpus of international studies devoted to the prediction of judicial decisions and the prospects of using AI in the work of professional judges.

Currently, no studies have modeled the decisions of Russian judges based on any one of the popular deep learning algorithms. The available studies on the factors of judicial decision-making (e. g., [Volkov, 2016]) use regression models in their analysis. The *Algorithm of Light* project [9], as well as the study [Zhuchkova, Kazun, 2023] used machine learning to extract information about partnership or kinship relations between a murder defendant and the victim from the texts of sentences, which is probably the most relevant example related to AI technologies at the moment. However, the main analysis in the latter study was constructed using regression models. Thus, this systematic review may be the starting point for the emergence of the first domestic studies that predict judicial decisions.

A sampling of sources allows us to answer the following key questions: First, what exactly can AI predict and what questions is it capable of answering? Second, what empirical material (decisions of which courts and in which countries) was used as a basis for training the models? Third, what methods and approaches do the researchers use to obtain their results? Although the third question is not the key one for this study, we will give a brief overview of the approaches that are used to predict

[9] URL: https://github.com/LanaLob/algorithm_sveta(Accessed 06.10.2024).

DOI: 10.14515/monitoring.2024.5.2580

outcomes of court cases. Fourth, we will evaluate the accuracy of AI-assisted judgment prediction and consider the limitations associated with evaluating the accuracy of algorithms. Finally, we will discuss the limitations of the use of AI, including the ethical issues associated with its implementation.

What can AI predict?

The first step was to define the terminology. There is a significant difference between how the terms «prediction» and «classification» are used in a broad context and how they are used in computer science, and thus in most research articles on AI-assisted judgment prediction. When social scientists talk about classification, it usually means sorting into types or categorizing them into groups. In most articles on the role of AI, classification refers to a broader class of phenomena that includes outcome prediction. Medvedeva, Wieling, and Vols [Medvedeva, Wieling, and Vols, 2022] provide a useful division of research on AI-assisted judgment prediction into three groups: outcome identification, text categorization, and judgment prediction. We discuss each of these types further below, but it is important that they can all be referred to as categorizations in the research literature.

The term *prediction* (or *forcasting*) has two meanings. If we have an array of already issued court decisions, we can train the algorithm on, say, 80 % of these cases, and test it on the remaining 20 %. In this case, the algorithm predicts the decisions of the judges with some accuracy. There is no mistake in calling this a prediction, but it has nothing to do with the outcomes of future cases or decisions that real judges have yet to make. Therefore, we are dealing with prediction in the «weak» sense of the term. Prediction in the «strong» sense means that our program is capable of predicting actual judicial decisions in the future. Most studies modeling judicial decisions [Medvedeva, Wieling, Vols, 2022] understand the prediction in the «weak» version of this term.

We will also speak of prediction mostly in the «weak» sense, using the term *forecasting* as a synonym. Forecasting always involves some probability and confidence intervals, but it should be noted that the accuracy of such predictions may not fully reflect the model's ability to predict future decisions in real cases. In general, any forecast, whether it is a prediction of inflation, election results, or a court decision, is based on data about the past and proceeds from the premise that in the future, the actions of people (including specialists) will be determined by the same set of factors as before. In reality, this is most often the case, which makes forecasts in general an important benchmark for decision-making, but there is always a risk of new factors emerging that break previous patterns and make forecasts based on past data less reliable.

As noted above, the entire corpus of literature on the classification of judgments can be divided into three categories. The first type is a search in the text of the document for the verdict or dispositive part. This task did not present any difficulties to a human lawyer. In this case, machine learning is applied for the initial processing of a large amount of data and saving time. In this task, the goal of researchers and programmers is to approximate the quality of a machine to the results that a human can produce. Studies that set themselves the task of identifying the outcome may

DOI: 10.14515/monitoring.2024.5.2580

achieve 99 % accuracy, which means not predicting the judge's decisions but merely being able to pick out the place where the decision is described. In what follows, we do not refer to studies of this type for predicting judicial decisions.

The second type of research categorizes legal texts. The task is to identify relevant circumstances of cases or their characteristics, such as the gender of the accused or victim and the circumstances of the crime. Although at first glance it seems that the second type of tasks is close to the first, the categories may not be based on explicit textual fragments, but on analyzing the whole document, for example, if the ideological orientation of a court is being evaluated [Shaikh, Sahu, Anand, 2020]. In the case of extra-legal factor extraction, AI can already compete with humans, as humans cannot always categorize with 100 % accuracy. However, not every such task applies to predicting judicial decisions, but only to those related to categorizing decisions — whether a verdict is passed or not, whether a complaint is approved, and so on.

Finally, the third type of task is the actual prediction of the outcome — whether the verdict will be acquittal or convicting, whether the court will recognize the violation of human rights, and so on. In addition, one can predict the sentence length (imprisonment) or fine amount. There is a thin line that distinguishes between the second and third types. Researchers [Medvedeva, Wieling, and Vols, 2022] refer to the third type only those few models [Medvedeva et al., 2021; Sharma et al., 2015; Waltl et al., 2017] that predict judges' decisions based on data available before the judgment (i. e., such models are close to the «strong» definition of the term prediction). Predictions in the «weak» sense, on the other hand, fall into the second category. Due to the small number of studies of the third type, we further refer to models predicting judges' decisions as those falling into both the second and third groups.

What court decisions is AI already predicting?

Most studies on the predictions of judgments are based on data from the European Court of Human Rights (hereinafter — ECHR) [Chalkidis, Androutsopoulos, Aletras, 2019; O'Sullivan, Beel, 2019; Medvedeva, Vols, Wieling, 2020; Aletras et al., 2016; Kaur et al., 2019; Medvedeva et al., 2020; Quemy, Wrembel, 2020] and the U. S. Supreme Court [Kaufman, Kraft, Sen, 2019; Katz, Bommarito, Blackman, 2017; Sharma et al., 2015]. The judgements databases of these courts are publicly available, allowing new algorithms to be tested on them. The remaining research also tends to focus on the decisions of higher courts — the constitutional court of Turkey [Sert, Yıldırım, Haşlak, 2022], the supreme court of the Philippines [Virtucio et al., 2018], Taiwan [Kowsrihawat, Vateekul, Boonkwan, 2018], Brazil [Freitas et al., 2022], and India [Malik at al., 2021].

Significantly less research is done on large samples of general courts or issue-specific courts from around the world (e. g., on taxes [Alarie, Niblett, Yoon, 2017; Waltl et al., 2017], on commercial disputes [Bagherian-Marandi, Ravanshadnia, Akbarzadeh-T, 2021]). Recently, new open data has been released [Cui et al., 2023]. The largest and best known public database is Cail2018 [Zhong et al., 2020], which includes 2.6 million cases from the Chinese Supreme Court. It should be noted that few studies (e. g., [Jacob de Menezes-Neto, Clementino, 2022]) work with lower court data.

DOI: 10.14515/monitoring.2024.5.2580

Open databases ([Alali et al., 2021; Xiao et al., 2018; Malik et al., 2021; Sebők, Kiss, and Járay, 2023]) allow testing new algorithms for analyzing court cases, which in turn serves as a benchmark for researchers on the quality of AI models. To prove that the new model is better, it is sufficient to run tests on one of the available datasets and outperform the last best result.

How does AI predict the decisions?

It is beyond the scope of this review to describe the data methodology in detail, as the methods used to analyze court cases are numerous, varied, and require specific skills to master. Below, we will briefly describe the main types of approaches used in the research, which will allow us to better understand the opportunities and limitations associated with them.

Again, a few words should be said about terminology [Sert, Yıldırım, Haşlak, 2022: 8]. AI, machine learning, and deep learning can be correlated using the metaphor of a nesting doll because they are «nested» in each other. AI is the broadest concept and includes many methods of data analysis. Natural Language Processing (NLP), which allows a computer to «read» and «understand» human language, is one of the key uses of AI, but by no means the only one. In the case of judgment prediction, it is NLP that we are dealing with. The next level is machine learning, which is an integral part of AI. Deep learning is a variant of machine learning. When using any of these methods, texts of court decisions usually require preprocessing. In turn, data preprocessing can be based on different machine learning methods and have different degrees of accuracy.

The main distinction between case prediction methods is deep learning and other, generally simpler, machine learning methods [Alcántara Francia, Nunez-del-Prado, Alatrista-Salas, 2022]. Supervised learning is one of the standard types of machine learning. The idea of supervised learning is that the algorithm is trained on a sample of cases containing parameters related to outcomes, that is, judges' decisions. The quality of the algorithm's learning is tested on another sample, similar in structure, but containing only case parameters, and the algorithm should predict the result. The accuracy of the prediction in this case is equal to the percentage of correctly predicted decisions, which we discuss in more detail in the next section. Traditional forms of machine learning such as logistic regression or decision trees lend themselves well to interpretation, although some more complex models (e.g., random forest) are «black boxes» that are more difficult to interpret.

Deep learning methods use neural networks with many layers, which allows for the training of models on large datasets. Such models can work with millions of parameters, and it is impossible to select them manually. Therefore, they differ from other machine learning methods in terms of the complexity of models, large amounts of data, and the use of significant computational resources. The results of deep learning, unlike those of traditional machine learning methods, are difficult or even impossible to interpret, which creates serious limitations, as interpretability is crucial in the legal field.

Currently, pre-trained language models [Song et al., 2022], such as BERT [Devlin et al., 2019], which are deep learning models, are most commonly used to clas-

DOI: 10.14515/monitoring.2024.5.2580

sify legal texts and predict judicial decisions. This method assumes that the model is first trained on a large number of texts and then used in a variety of novel tasks. This approach differs significantly from earlier methods [Ashley, Brüninghaus, 2009], in which the algorithm was created specifically for a particular task and dealing with specific datasets. Algorithms such as BERT are more versatile but also limited because they may undertrain the specificity of legal texts. However, they can also be trained, and a specialized version can be created for specific tasks, such as LegalBERT trained on legal texts [Chalkidis et al., 2020]. Studies have shown that specialized language models can produce 1—5 % higher accuracy rates than general models [Song et al., 2022], but their training is associated with additional costs for the researcher. The mentioned BERT and LegalBERT are only two current examples of pre-trained language models, but there are many others. It is beyond the scope of this study to analyze in more depth the features of the different models, which could be found in [Alcántara Francia, Nunez-del-Prado, Alatrista-Salas, 2022; Song et al., 2022].

Note that the generative artificial intelligence technology ChatGPT also belongs to the class of pre-trained language models, although the main task of ChatGPT is text generation rather than classification. Therefore, BERT and similar models are better suited to legal text classification and judgment prediction than ChatGPT, which, in turn, is better suited for legal document generation.

With what accuracy can you predict a court's decision?

As noted above, the classification of outcomes ranges from accurately identifying the dispositive part of a sentence to predicting actual judgments that have not yet been rendered. At the same time, the outcomes themselves can vary: guilty or not guilty, plaintiff's complaint accepted or rejected [Bagherian-Marandi et al., 2021], appeal successful or not [Jacob de Menezes-Neto, Clementino, 2022; Waltl et al., 2017], rights recognized as violated or not [Sert, Yıldırım, Haşlak, 2022], etc. Generally, prediction accuracy is assessed for binary variables, although the size of the fines or sentence lengths, albeit with less accuracy, can also be predicted.

Researchers have identified several key metrics for calculating accuracy [ibid:] *Accuracy*, *Precision*, *Recall*, and *F1 Score*. *Accuracy* is the sum of correctly categorized positive and negative decisions divided by their total number. *Precision* and *Recall* refer to the proportion of correctly predicted negative decisions and correctly predicted positive decisions, respectively. *The F1 Score* is often used to compare the quality of different models. This metric is calculated as follows:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Recision + Recall}.$$

When comparing the accuracy of different studies, it is necessary to consider which of the parameters the authors derive as *Accuracy* or *F1 Score*, as this can yield differences from a few tenths of a percent to a few percent. Although most studies have used these indices, they have recently been criticized for their overly optimistic results [Chicco, Jurman, 2020]. As a replacement, the *Matthews correlation coeffi-*

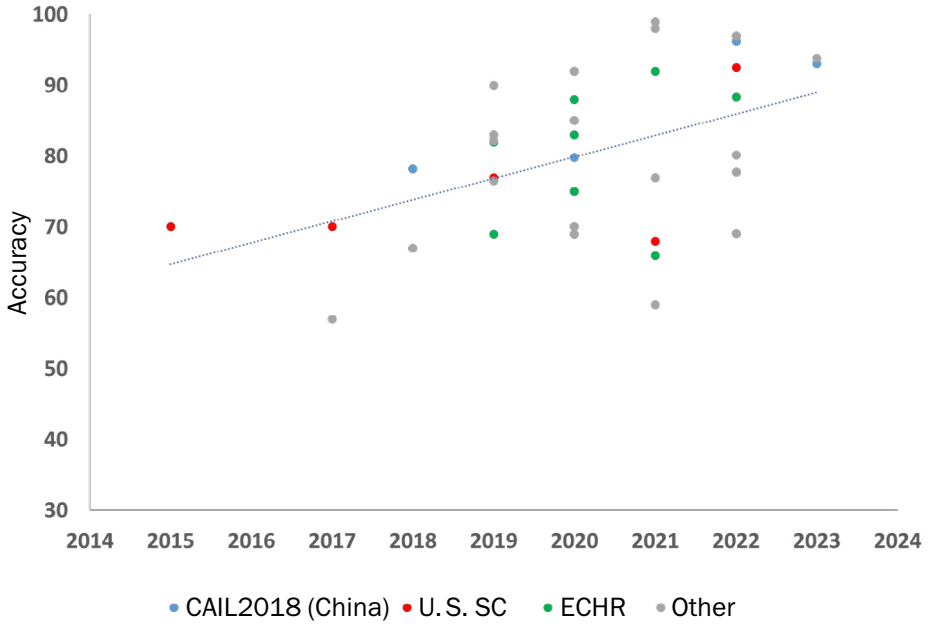
DOI: 10.14515/monitoring.2024.5.2580

cient (MCC) is proposed, which provides good results only if the algorithm performs well in all four variants: true positive, true negative, false positive, and false negative.

The accuracy of the algorithms typically ranges from 60 % to 99 %, but a lower accuracy rate does not necessarily mean that the algorithm is bad; it might be good, just that it was used to solve a more difficult problem. For this reason, the descriptive analysis we offer below should be taken precisely as an illustration of the underlying trend and variation in accuracy in predicting the outcome of court cases in existing works.

The analysis of the international experience based on the 34 selected articles allowed us to assess the dynamics of the predictive power of the models classifying the texts of judicial decisions (see Fig. 1).

Figure 1. Dynamics of accuracy of models predicting court decisions from 2015 to 2023 (for different databases)[10]



The resulting picture reflects a general increasing trend in the predictive power of the models. However, the graph also offers a simplified picture, because different studies have different objectives, analysis algorithms, and accuracy estimates. For example, the study [Jacob de Menezes-Neto, Clementino, 2022] is not in the graph

[10] Note: The graph was constructed by the author based on 38 assessments made within the framework of 34 studies. Of these, 19 observations are readily available from the systematic review [Medvedeva, Wieling, Vols, 2022], the rest were collected and coded by the author. The sample includes only estimates for court decisions involving decisions with two outcomes (1 or 0). Vertical axis — *Accuracy*; for some studies lacking an *Accuracy* value, an *F1* Score is provided; horizontal axis — year of publication of the study. The blue line on the graph is the trend line for all observations.

DOI: 10.14515/monitoring.2024.5.2580

because it used an alternative method of accuracy assessment that is not directly comparable to other studies.

From a methodological perspective, it is better to compare the accuracy of solving similar problems performed on the same data, such as the database of ECHR decisions, the U. S. Supreme Court, or the Cail2018 database. These popular open datasets were published, along with benchmarks that set a baseline level of quality for all future algorithms. If we compare progress within individual databases, we also observe an increasing trend. For example, the prediction accuracy of the Chinese database Cail2018 has increased from 78 % [Xiao et al., 2018] in 2018 to 96 % in 2023 [Cui et al., 2023]. In 2015—2017, algorithms predicted U. S. Supreme Court decisions with 70 % accuracy [Katz, Bommarito, and Blackman, 2017; Sharma et al., 2015], and in 2022, there are models that predict them with 92.5 % accuracy [Alghazzawi et al., 2022]. Algorithms modeling ECtHR decisions also predict decisions with an accuracy of 92 % [Medvedeva et al., 2020]. However, as noted earlier, we are talking about prediction in its «weak» form, that is, the quality of the model trained on already available decisions and tested on the same data. These models will provide an adequate forecast of the future only if external circumstances, such as legislation, the composition of the court, and the political environment, do not change significantly.

Another interesting question regarding the accuracy of predictions is related to the choice of the point of comparison. The natural choice seems to be to compare the model with real decisions. This approach was used in all studies included in Figure 1. However, there is an alternative point of view: to understand whether AI can replace a lawyer, it is more appropriate to compare the predictive abilities of a real person who is an expert in a particular field. For example, in [Jacob de Menezes-Neto, Clementino, 2022], court data on more than 765,000 cases from Brazil were used to compare the accuracy of predicting the outcome of appeals by an AI or 22 experts from among judges and court staff (who evaluated a random sample of 690 cases). According to the results, AI outperformed those of humans by almost three times. However, this is not the first study to demonstrate that AI can outperform humans. In a study [Ruger et al., 2004], the machine also outperformed legal experts in predicting the decisions of the U. S. Supreme Court, but not by such a high margin.

Several conclusions can be drawn from the above. First, the accuracy of algorithms for classifying judicial decisions strongly depends on the data and on the tasks set for researchers. Second, in recent years there has been a clear trend towards improving the quality of predictive models. When compared with the predictive abilities of legal experts, some algorithms show clear superiority over humans. However, we cannot yet say of a hundred percent accuracy that would allow us to replace a judge.

Debate: Can AI replace a judge?

Before answering our research question, we summarize the main limitations of the use of AI in jurisprudence.

DOI: 10.14515/monitoring.2024.5.2580

First, the results of the models were highly dependent on data quality. For a model to make good predictions, the texts of court decisions should have a similar structure, and there should be a sufficient number of such texts for training. Therefore, unique decisions of constitutional courts cannot be predicted by existing analysis methods [Sert, Yıldırım, Haşlak, 2022]. Second, a separate technical challenge is related to working in languages other than English, such as Chinese, Turkish [ibid.], or Portuguese [Jacob de Menezes-Neto, Clementino, 2022]. However, this limitation is surmountable —pretrained language models can be further trained. Third, algorithms built on small or specific samples have less practical utility than those built for general courts [ibid.]. Algorithms have learned to predict accurately (in the «weak» sense of the term) the decisions of the U. S. Supreme Court and the ECHR, but this does not allow the results to be extended to lower-level courts. The issue of developing models that predict court decisions in typical cases that affect the mainstream population has not yet been resolved. The available research in this area is extremely sparse, and there are no universal solutions. Fourth, most models deal with predictions in the «weak» sense of the term and do not guarantee that one can accurately predict future judicial decisions. We can describe this as a problem of induction; no matter how many observations we make about the past, they do not allow us to look into the future. Thus, the high accuracy of AI predictions in Figure 1 should not mislead us: it merely reflects the good training of the models on a very limited amount of past data.

Finally, a key problem with deep learning models is that they are difficult to interpret. A human judge can explain the logic behind his or her decision. Classical models such as regressions and decision trees also produce easily interpretable results. For example, from the regression model on the influence of various factors on the decisions of Russian judges in murder cases [Zhuchkova, Kazun, 2023], it is easy to understand how each analyzed parameter of the case (gender, confession of guilt, re-offense, etc.) affect the verdict. However, judgment prediction methods based on deep learning are «black boxes». While these models make very accurate predictions, they do not allow researchers to understand the legal and extra-legal factors behind each specific decision. The advantage of AI over linear regression models is that the latter can underestimate the nonlinearity of the influence of many factors [Alarie et al., 2017]. In addition, models based on deep learning provided the most accurate predictions.

One of the most highly cited papers on machine learning-based decision-making [Rudin, 2019] contains a key thesis in its title, which is — one should not use «black boxes» for decision-making when the costs of error are high. In this paper, the author calls the classic dilemma between the explanatory power of a model and its interpretability a myth. Interpretable models also have high predictive power. For example, in a study [Tan et al., 2020], an interpretable model was built on 1.3 million crime cases in China with an accuracy comparable to or better than the results of black-box models. A key advantage of interpretable models is their ability to handle errors. If an AI algorithm makes an incorrect judgment (or suggests the wrong course of treatment), it can have a significant impact on a person's fate. If a black box makes a similar mistake, we will not be able to tell why it was made or wheth-

DOI: 10.14515/monitoring.2024.5.2580

er it will occur again in the future. Owing to this shortcoming, such systems may never gain a high level of public trust. However, this does not mean that the development of black box prediction in medicine or law should stop; if such algorithms have statistically higher accuracy, for some tasks, accuracy and speed may outweigh transparency.

In 2016, the Wisconsin Supreme Court ruled that judges could rely on the results of the analysis performed by algorithms when making decisions, even if the algorithm's working principle was not completely transparent [Beriain, 2018]. The reasoning behind this decision was summarized in two theses: the quality of algorithms is very high, and the judge is competent enough to make an informed decision on whether to rely on the results of an algorithm. In fact, this precedent has paved the way for the use of AI as a judge's assistant in decision-making in the US. The reason for this trial was the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system. According to its creators, COMPAS is supposed to predict the individual risk of recidivism, which is an important factor when making decisions. However, the court's decision sparked debate about the accuracy of the algorithm [Dressel, Farid, 2018] and whether it reproduces racial stereotypes drawn from learning from people's decisions. Be that as it may, without this system, judges must still assess the risks that a defendant will reoffend, and this process is not free from various biases [Brennan, Dieterich, 2018].

In general, the idea that a significant proportion of judicial verdicts can be predicted should change the idea of what constitutes judicial discretion [Tahura, Selvadurai, 2023]. However, not all legal decisions have a single solution. If an AI algorithm is going to be able to predict judges' decisions very accurately, it means that it must reproduce not only a formal analysis of legal and extra-legal factors, but also what judges themselves have called the «psychology of litigation» [11]. Without the development of general AI, which has not been discussed in this article, the algorithm is unlikely to be able to understand the nuances of language, wordplay, values, and meanings perceived by people in different case circumstances (but even a «weak» AI might well learn over time to mimic this understanding if trained on a sufficient volume of such cases). Hence, there is an important concern about whether AI reproduces judges' stereotypes, ideology, and other beliefs in its analysis [Manresa-Yee, Ramis, 2022].

In this context, an even more interesting question arises: should AI fully predict the decisions of a human judge, or can it go a step higher in this respect? Research shows [Doerner, 2015; Franklin, Fearn, 2008] that judges in the US are not free from racial and gender stereotypes. Moreover, under conditions of bounded rationality [Albonetti, 1991] associated with information and time constraints, it is these stereotypes that often become reference points [Steffensmeier, Ulmer, Kramer, 1998] that help make difficult decisions. After all, an AI can follow the letter of the law more strictly than a human (not to mention a higher speed of analyzing information). Humans are not free from sympathies, values, beliefs, and psychological pressures, while AI follows these factors only if it has been taught to follow them. This is a tricky

[11] The Council of Judges rejected the idea of replacing a judge with an AI // Pravo.ru. 2023. October 26. URL: https://pravo.ru/news/249529/(Accessed 06.10.2024).

DOI: 10.14515/monitoring.2024.5.2580

question that has not been answered yet. However, we can predict that at least the population will not be ready for such a turn of events for a long time.

The study [Barysė, Sarel, 2023] puts forward an important thesis that the legitimacy of the use of AI is perceived differently by the public and professionals depending on the stage of the judicial process where the technology is used. People generally trust human judges more than AI; therefore, they do not support high court automation at the decision-making stage. However, the public favors the use of technology at the data collection stage because it can help improve the objectivity of human analysis.

Conclusion: AI as a judge's assistant

The results of this study suggest several practical recommendations for the development of AI technologies to assist lawyers and judges. Although AI cannot yet replace a judge, it can assist judges in their professional activities.

We have shown that despite the rapid improvement in algorithms that predict judicial decisions, most of them suffer from data-related limitations. For example, algorithms designed to analyze the decisions of higher courts have low external validity. Most citizen and business litigations occur in lower-level courts, which AI is not yet adept at predicting. Except for a single dataset from China [Xiao et al., 2018], other publicly available data include an extremely limited number of cases, not exceeding a few tens of thousands. There are already commercial projects in Russia that can predict the outcomes of arbitration cases, but their algorithms are opaque because they have not been published in peer-reviewed academic journals.

In Russia, according to the law of December 22, 2008 № 262-FZ «On Ensuring Access to Information on the Activity of Courts in the Russian Federation», a huge amount of data on criminal, civil and arbitration cases is published, which creates a great potential for AI to learn from them. This means that although AI models that would predict Russian court decisions have not been created to date, significant breakthroughs may be made in this area in the foreseeable future. Research on using AI to classify legal texts worldwide is still at the beginning of its development.

Another question is whether AI technology should be introduced into judicial processes. It is, of course, impossible to answer this question unequivocally, and the debate on this issue demonstrates the complexity of the ethical issues involved. However, the complexity of ethical issues does not negate the importance of technology development itself. One promising possibility is to turn AI into an assistant that can assist judges in working with case files, including categorizing them, finding problem areas, and highlighting the most significant circumstances in a case. The public position of one of the judges of the Supreme Court of the Russian Federation [12] generally supports this very scenario.

It is important to keep in mind that most of the existing algorithms for predicting court decisions operate on a black box principle, which severely limits the potential for their implementation in the legal profession. Algorithms whose operations are not understood cannot be thoughtlessly integrated into high-stakes settings.

[12] Burnov V. VS: AI can help reduce the number of errors in sentencing // RAPSI. 2024. July 1. URL: https://www.rapsinews.ru/judicial_mm/20240701/310059533.html(Accessed 05.07.2024).

DOI: 10.14515/monitoring.2024.5.2580

However, the dilemma between the accuracy of decisions and transparency of the decision-making process may be false [Rudin, 2019]. In such a case, judicial practice should precisely introduce solutions that allow not only to predict decisions but also to explain them. The latter task should be on the agenda of domestic research in the field of using AI technologies in the judicial process.

DOI: 10.14515/monitoring.2024.5.2580

References

1. Alali M., Syed S., Alsayed M., Patel S., Bodala H. (2021) JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction. *arXiv*. Art. 2112.03414. https://doi.org/10.48550/arXiv.2112.03414.

2. Alarie B., Niblett A., Yoon A. (2017) Using Machine Learning to Predict Outcomes in Tax Law. *SSRN*. https://doi.org/10.2139/ssrn.2855977.

3. Albonetti C. A. (1991) An Integration of Theories to Explain Judicial Discretion. *Social Problems*. Vol. 38. No. 2. P. 247—266. https://doi.org/10.2307/800532.

4. Alcántara Francia O. A., Nunez-del-Prado M., Alatrista-Salas H. (2022) Survey of Text Mining Techniques Applied to Judicial Decisions Prediction. *Applied Sciences*. Vol. 12. No. 20. Art. 20. https://doi.org/10.3390/app122010200.

5. Aletras N., Tsarapatsanis D., Preoţiuc-Pietro D., Lampos V. (2016) Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science*. Vol. 2. No. 10. Art. e93. https://doi.org/10.7717/peerj-cs.93.

6. Alghazzawi D., Bamasag O., Albeshri A., Sana I., Ullah H., Asghar M. Z. (2022) Efficient Prediction of Court Judgments Using an LSTM+CNN Neural Network Model with an Optimal Feature Set. *Mathematics*. Vol. 10. No. 5. Art. 5. https://doi.org/10.3390/math10050683.

7. Ashley K. D., Brüninghaus S. (2009) Automatically Classifying Case Texts and Predicting Outcomes. *Artificial Intelligence and Law*. Vol. 17. No. 2. P. 125—165. https://doi.org/10.1007/s10506-009-9077-9.

8. Bagherian-Marandi N., Ravanshadnia M., Akbarzadeh-T M.-R. (2021) Two-Layered Fuzzy Logic-Based Model for Predicting Court Decisions in Construction Contract Disputes. *Artificial Intelligence and Law*. Vol. 29. No. 4. P. 453—484. https://doi.org/10.1007/s10506-021-09281-9.

9. Barysė D., Sarel R. (2023) Algorithms in the Court: Does It Matter Which Part of the Judicial Decision-Making is Automated? *Artificial Intelligence and Law*. Vol. 32. P. 117–146. https://doi.org/10.1007/s10506-022-09343-6.

10. Beriain I. D. M. (2018) Does the Use of Risk Assessments in Sentences Respect the Right to Due Process? A Critical Analysis of the *Wisconsin* v. *Loomis Ruling*. *Law, Probability & Risk*. Vol. 17. No. 1. P. 45–53. https://doi.org/10.1093/lpr/mgy001.

11. Brennan, T., Dieterich W. (2018) Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In: Singh J. P., Kroner D. G., Wormith J. S., Desmarais S. L., Hamilton Z. (eds.) *Handbook of Recidivism Risk/Needs Assess-*

ment Tools. Hoboken: John Wiley & Sons. P. 49—75. https://doi.org/10.1002/9781119184256.ch3.

12. Chalkidis I., Androutsopoulos I., Aletras N. (2019) Neural Legal Judgment Prediction in English. *arXiv*. Art. 1906.02059. https://doi.org/10.48550/arXiv.1906.02059.

13. Chalkidis I., Fergadiotis M., Malakasiotis P., Aletras N., Androutsopoulos I., Androutsopoulos I. (2020) LEGAL-BERT: The Muppets Straight Out of Law School. *arXiv*. Art. 2010.02559. https://arxiv.org/abs/2010.02559.

14. Chicco D., Jurman G. (2020) The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*. Vol. 21. no. 1. Art. 6. https://doi.org/10.1186/s12864-019-6413-7.

15. Devlin J., Chang M.-W., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J., Doran C., Solorio T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. *Vol. 1.* Minneapolis, Association for Computational Linguistics. P. 4171—4186. https://doi.org/10.18653/v1/N19-1423.

16. Doerner J. K. (2015) The Joint Effects of Gender and Race/Ethnicity on Sentencing Outcomes in Federal Courts. *Women & Criminal Justice*. Vol. 25. No. 5. P. 313—338. https://doi.org/10.1080/08974454.2014.989298.

17. Franklin C. A., Fearn N. E. (2008) Gender, Race, and Formal Court Decision-Making Outcomes: Chivalry/Paternalism, Conflict Theory, or Gender Conflict? *Journal of Criminal Justice*. Vol. 36. No. 3. P. 279—290. https://doi.org/10.1016/j.jcrimjus.2008.04.009.

18. Freitas A. L., Allende-Cid H., Santana O., Oliveira-Lage L. (2022) Predicting Brazilian Court Decisions. *PeerJ Computer Science*. Vol. 8. Art. e904. https://doi.org/10.7717/peerj-cs.904.

19. Gerber M. S. (2014) Predicting Crime Using Twitter and Kernel Density Estimation. *Decision Support Systems*. Vol. 61. P. 115—125. https://doi.org/10.1016/j.dss.2014.02.003.

20. Jacob de Menezes-Neto E., Clementino M. B. M. (2022) Using Deep Learning to Predict Outcomes of Legal Appeals Better than Human Experts: A Study with Data from Brazilian Federal Courts. *PLoS ONE*. Vol. 17. No. 7. Art. e0272287. https://doi.org/10.1371/journal.pone.0272287.

21. Dressel J., Farid H. (2018) The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*. Vol. 4. No. 1. Art. eaao5580. https://doi.org/10.1126/sciadv.aao5580.

DOI: 10.14515/monitoring.2024.5.2580

22. Cui J., Shen X., Wen S. (2023) A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges. *IEEE Access*. Vol. 11. P. 102050—102071. https://doi.org/10.1109/access.2023.3317083.

23. Katz D. M., Bommarito M. J., Blackman J. (2017) A General Approach for Predicting the Behavior of the Supreme Court of the United States. *PLoS ONE*. Vol. 12. no. 4. Art. e0174698. https://doi.org/10.1371/journal.pone.0174698.

24. Kaufman A. R., Kraft P., Sen M. (2019) Improving Supreme Court Forecasting Using Boosted Decision Trees. *Political Analysis*. Vol. 27. no. 3. P. 381–387. https://doi.org/10.1017/pan.2018.59.

25. Kaur H., Choudhury T., Singh T. P., Shamoon Mohammad. (2019) Crime Analysis using Text Mining. In: Ming Fong A.Ch., Hong G. Y., Fong B. (eds.) *2019 International Conference on Contemporary Computing and Informatics (IC3I)*. Singapore. P. 283–288. https://doi.org/10.1109/IC3I46837.2019.9055606.

26. Kort F. (1957) Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the "Right to Counsel" Cases. *American Political Science Review*. Vol. 51. No. 1. P. 1–12. https://doi.org/10.2307/1951767.

27. Kovalenko K. E., Pechatnova Yu. V., Statsenko D. A., Kovalenko N. E. (2020) The Robot Judge as a Resolution of Judicial Discretion Contradictions (Legal Aspect). *Legal Bulletin of Dagestan State University*. Vol. 36. No. 4. P. 169—173. (In Russ.)
Коваленко К. Е., Печатнова Ю. В., Стаценко Д. А., Коваленко Н. Е. Судья-робот как преодоление противоречий судебного усмотрения (юридический аспект) // Юридический вестник Дагестанского Государственного Университета. 2020. Т. 36. № 4. С. 169—173.

28. Kowsrihawat K., Vateekul P., Boonkwan P. (2018) Predicting Judicial Decisions of Criminal Cases from Thai Supreme Court Using Bi-directional GRU with Attention Mechanism. In: Do Van T., Do Duc H., Kravchuk N. V. (2021) Artificial Intelligence as a Judge: Prospects and Concerns (Review). *Social and Human Sciences. Domestic and Foreign Literature. Series 4: State and Law*. No. 1. P. 115—122. (In Russ.)
Кравчук Н. В. Искусственный интеллект как судья: перспективы и опасения (Обзор) // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 4: Государство и право. 2021. № 1. С. 115—122.

29. Nguyen G. (eds.) *5th Asian Conference on Defense Technology (ACDT)*. Hanoi, Vietnam. P. 50–55. https://doi.org/10.1109/acdt.2018.8592948.

30. Malik V., Sanjay R., Nigam S. K., Ghosh K., Guha S. K., Bhattacharya A., Bhattacharya A., Bhattacharya A., Modi A. (2021) ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In: Zong Ch., Xia F., Li W., Navigli R. (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Confer*. Strouds-

DOI: 10.14515/monitoring.2024.5.2580

burg: Association for Computational Linguistics. P. 4046—4062. https://doi.org/10.18653/v1/2021.acl-long.313.

31. Manresa-Yee C., Ramis S. (2022) Assessing Gender Bias in Predictive Algorithms using eXplainable AI *arXiv*. Art. 2203.10264. https://doi.org/10.48550/arXiv.2203.10264.

32. Medvedeva M., Üstun A., Xu X., Vols M., Wieling M. (2021) Automatic Judgement Forecasting for Pending Applications of the European Court of Human Rights. In: Ashley K. D., Atkinson K., Branting K., Francesconi E., Grabmair M., Walker V. R., Waltl B., Wyner A. Z. (eds.) Proceedings of the European Court of Human Rights. (eds.) *Proceedings of the fifth workshop on automated semantic analysis of information in legal text (ASAIL 2021), São Paulo, Brazil*. CEUR Workshop Proceedings. P. 1—12. https://ceur-ws.org/Vol-2888/paper2.pdf.

33. Medvedeva M., Vols M., Wieling M. (2020) Using Machine Learning to Predict Decisions of the European Court of Human Rights. *Artificial Intelligence and Law*. Vol. 28. No. 2. P. 237—266. https://doi.org/10.1007/s10506-019-09255-y.

34. Medvedeva M., Wieling M., Vols M. (2022) Rethinking the Field of Automatic Prediction of Court Decisions. *Artificial Intelligence and Law*. Vol. 31. P. 195—212. https://doi.org/10.1007/s10506-021-09306-3.

35. Medvedeva M., Xiao X., Wieling M., Vols M. (2020) JURI SAYS: An Automatic Judgement Prediction System for the European Court of Human Rights. *Frontiers in Artificial Intelligence and Applications*. Vol. 334. P. 277–280. https://doi.org/10.3233/faia200883.

36. O'Sullivan C., Beel J. (2019) Predicting the Outcome of Judicial Decisions made by the European Court of Human Rights. *arXiv*. Art. 1912.10819. https://doi.org/10.48550/arXiv.1912.10819.

37. Quemy A., Wrembel R. (2020) On Integrating and Classifying Legal Text Documents. In: Hartmann S., Küng J., Kotsis G., Tjoa A. M., Khalil I. (eds.) *Database and Expert Systems Applications*. Cham, Switzerland: Springer International Publishing. P. 385–399. https://doi.org/10.1007/978-3-030-59003-1_25.

38. Rudin C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*. Vol. 1. No. 5. Art. 5. https://doi.org/10.1038/s42256-019-0048-x.

39. Ruger T. W., Kim P. T., Martin A., Martin A. D., Quinn K. M. (2004) The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking. *Columbia Law Review*. Vol. 104. No. 4. P. 1150–1209. https://doi.org/10.2307/4099370.

40. Rusakova E. P. (2021) Integration of “Smart” Technologies in the Civil Proceedings of the People’s Republic of China. *RUDN Journal of Law*. Vol. 25. No. 3. P. 622–633. https://doi.org/10.22363/2313-2337-2021-25-3-622-633.

41. Sebők M., Kiss R., Járay I. (2023) Introducing HUNCOURT: A New Open Legal Database Covering the Decisions of the Hungarian Constitutional Court for Between 1990 and 2021. *Journal of the Knowledge Economy*. Vol. 15. P. 6507—6540. https://doi.org/10.1007/s13132-023-01395-6.

42. Sert M. F., Yıldırım E., Haşlak İ. (2022) Using Artificial Intelligence to Predict Decisions of the Turkish Constitutional Court. *Social Science Computer Review*. Vol. 40. No. 6. P. 1416—1435. https://doi.org/10.1177/08944393211010398.

43. Shaikh R. A., Sahu T. P., Anand V. (2020) Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers. *Procedia Computer Science*. Vol. 167. P. 2393–2402. https://doi.org/10.1016/j.procs.2020.03.292.

44. Sharma R. D., Mittal S., Tripathi S., Acharya S. (2015) Using Modern Neural Networks to Predict the Decisions of Supreme Court of the United States with State-of-the-Art Accuracy. In: Arik, S., Huang, T., Lai, W., Liu, Q. (eds.) *Neural Information Processing. ICONIP 2015. Lecture Notes in Computer Science. Vol. 9490*. Cham: Springer. P. 475—483. https://doi.org/10.1007/978-3-319-26535-3_54.

45. Shi C., Sourdin T., Li B. (2021) The Smart Court — A New Pathway to Justice in China? *International Journal of Court Administration.* Vol. 12. No. 1. Art. 1. https://doi.org/10.36745/ijca.367.

46. Song D., Gao S., He B., Schilder F. (2022) On the Effectiveness of Pre-Trained Language Models for Legal Natural Language Processing: An Empirical Study. *IEEE Access*. Vol. 10. P. 75835—75858. https://doi.org/10.1109/ACCESS.2022.3190408.

47. Sourdin T., Cornes R. (2018) Do Judges Need to Be Human? The Implications of Technology for Responsive Judging. In: Sourdin T., Zariski A. (eds.) *The Responsive Judge: International Perspectives*. Singapore: Springer. P. 87—119. https://doi.org/10.1007/978-981-13-1023-2_4.

48. Steffensmeier D., Ulmer J., Kramer J. (1998) The Interaction of Race, Gender, and Age in Criminal Sentencing: The Punishment Cost of Being Young, Black, and Male. *Criminology*. Vol. 36. No. 4. P. 763—798. https://doi.org/10.1111/j.1745-9125.1998.tb01265.x.

49. Stepanov O. A., Basangov D. A. (2022) On the Prospects of Artificial Intelligence Impact on Judiciary. *Bulletin of Tomsk State University. Philosophy, Sociology, Political Science*. No. 475. P. 229—237. (In Russ.)
Степанов О. А., Басангов Д. А. О перспективах влияния искусственного интеллекта на судопроизводство // Вестник Томского Государственного Университета. Философия. Социология. Политология. 2022. № 475. С. 229—237.

DOI: 10.14515/monitoring.2024.5.2580

50. Tahura U. S., Selvadurai N. (2023) The Use of Artificial Intelligence in Judicial Decisionmaking: The Example of China. *International Journal of Law, Ethics, and Technology*. Vol. 3. P. 1—20. https://doi.org/10.55574/pyeb5374.

51. Tan H., Zhang B., Zhang H., Li R. (2020) The Sentencing-Element-Aware Model for Explainable Term-of-Penalty Prediction. In: Zhu, X., Zhang, M., Hong, Y., He, R. (eds.) *Natural Language Processing and Chinese Computing. NLPCC 2020. Lecture Notes in Computer Science. Vol. 12431.* Cham: Springer. P. 16–27. https://doi.org/10.1007/978-3-030-60457-8_2.

52. Taylor I. (2023) Justice by Algorithm: The Limits of AI in Criminal Sentencing. *Criminal Justice Ethics*. Vol. 42. No. 3. P. 193–213. https://doi.org/10.1080/0731129X.2023.2275967.

53. Virtucio M. B. L., Aborot J. A., Abonita J. K. C., Avinante R. S., Copino R. J. B., Neverida M. P., Osiana V. O., Peramo E. C., Syjuco J. G., Tan G. B. A. (2018) Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning. In: Honiden S., Fujii R. (eds.) *2018 IEEE 42nd Annual IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. Tokyo. P. 130–135. https://doi.org/10.1109/compsac.2018.10348.

54. Volkov V. (2016) Legal and Extralegal Origins of Sentencing Disparities: Evidence from Russia's Criminal Courts. *Journal of Empirical Legal Studies*. Vol. 13. No. 4. P. 637—665. https://doi.org/10.1111/jels.12128.

55. Waltl B., Bonczek G., Scepankova E., Landthaler J., Matthes F. (2017) Predicting the Outcome of Appeal Decisions in Germany's Tax Law. Electronic Participation. In: Parycek P. et al. (eds.) *Electronic Participation. ePart 2017. Lecture Notes in Computer Science. Vol. 10429.* Cham: Springer. P. 89—99. https://doi.org/10.1007/978-3-319-64322-9_8.

56. Xiao C., Zhong H., Guo Z., Tu C., Liu Z., Sun M., Feng Y., Han X., Hu Z., Wang H., Xu J. (2018) CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv*. Art. 1807.02478. https://arxiv.org/abs/1807.02478.

57. Xu Z. (2022) Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities. *Applied Artificial Intelligence*. Vol. 36. No. 1. Art. 2013652. https://doi.org/10.1080/08839514.2021.2013652.

58. Zakharcev S. I., Salnikov V. P. (2018) The Robot Judge in Criminal Proceedings: Good or Bad? *Legal Science: History and Modernity*. No. 7. P. 176—180. (In Russ.) Захарцев С. И., Сальников В. П. Судья-робот в уголовном процессе: хорошо или плохо? // Юридическая наука; история и современность. 2018. № 7. С. 176—180.

59. Zhong H., Xiao C., Tu C., Tianyang Zhang, Zhiyuan L., Liu Z., Liu Z., Sun M. (2020) How Does NLP Benefit Legal Systems: A Summary of Legal Artificial Intelligence. In: Jurafsky D., Chai J., Schluter N., Tetreault J. (eds.) *Proceedings of the 58th An-*

DOI: 10.14515/monitoring.2024.5.2580

nual Meeting of the Association for Computational Linguistics. Stroudsburg. Association for Computational Linguistics. P. 5218–5230. https://aclanthology.org/2020.acl-main.466/.

60. Zhuchkova S., Kazun A. (2023) Exploring Gender Bias in Homicide Sentencing: An Empirical Study of Russian Court Decisions Using Text Mining. *Homicide Studies*. https://doi.org/10.1177/10887679231217159.