Artificial Intelligence and Legal Decision-Making: The Wide Open?

A Study Examining International Arbitration

Maxi SCHERER*

The article explores the use of Artificial Intelligence (AI) in arbitral or judicial decision-making from a holistic point of view, exploring the technical aspects of AI, its practical limitations as well as its methodological and theoretical implications for decision-making as a whole. While this article takes the angle of international arbitration, it looks at examples and studies from a wide variety of legal areas and its conclusions are relevant for adjudicatory decision-making more globally. The author assesses existing studies on decision outcome prediction and concludes that the methodology and assumptions employed put into doubt the claim these models might be used for ex ante outcome predictions. The article also discusses whether AI models, which are typically based on information extracted from previous input data, are likely to follow 'conservative' approaches and might not be adapted to deal with important policy changes over time. The article further finds that a blind deferential attitude towards algorithmic objectivity and infallibility is misplaced and that AI models might perpetuate existing biases. It discusses the need for reasoned decisions, which is likely to be an important barrier for AI-based legal decision-making. Finally, looking at existing legal theories on judicial decision-making, the article concludes that the use of AI and its reliance on probabilistic inferences could constitute a significant paradigm shift. In the view of the author, AI will no doubt fundamentally affect the legal profession, including judicial decision-making, but its implications need to be considered carefully.

1 INTRODUCTION

> L'avenir n'est jamais que du présent à mettre en ordre. Tu n'as pas à le prévoir, mais à le permettre. *Antoine de Saint-Exupéry*

The relationship of the future to the present is the topic of de Saint-Exupéry's somewhat mysterious quote. The first sentence states that the future simply is the present in better order or better organized. Concerning the future, the second sentence goes on, the task is not to foresee it, but to allow or enable it. How better

* Professor of Law, Queen Mary University of London; Special Counsel, Wilmer Cutler Pickering Hale and Dorr LLP. The author wishes to thank Jose Alvarez (NYU), Pierre Brunet (Université de Paris I Panthéon-Sorbonne), Jacques de Werra (Université de Genève), Susan Franck (American University, Washington College of Law), Mark Kantor (via OGEMID), Elizabeth Whitsitt (University of Calgary), for having provided most helpful comments and thoughts on an earlier version of the article. Email: maxi.scherer@wilmerhale.com.

Scherer, Maxi. 'Artificial Intelligence and Legal Decision-Making: The Wide Open?'. *Journal of International Arbitration* 36, no. 5 (2019): 539–574.

to enable a more organized future than with the use of technology, such as Artificial Intelligence (AI)? It is trite to underline the importance AI already has in our daily lives. Whether we are aware of it or not, AI is used to filter spam emails, write newspaper articles, provide medical diagnoses, and assess access to credits.

Nevertheless, lawyers typically believe that the impact on their profession will be limited. This ignores that AI already touches many areas of law, including contract analysis, legal research, e-discovery, etc.[1] For instance, computer programs are available to help lawyers to analyse the other side's written submissions and to provide relevant case law that was omitted therein or rendered since. Unsurprisingly, AI in law is a growing business.[2]

In international arbitration, the use of AI has been predicted for a wide variety of tasks, including appointment of arbitrators, legal research, drafting and proof-reading of written submissions, translation of documents, case management and document organization, cost estimations, hearing arrangements (such as transcripts or simultaneous foreign language interpretation), and drafting of standard sections of awards (such as procedural history).[3]

This article will not deal with those aspects but instead focus on one of the more controversial areas which is at the core of the arbitral process: the decision-making itself.[4]

---

1 *See* e.g. Richard Susskind, *Tomorrow's Lawyers: An Introduction to Your Future* (2d ed., Oxford University Press 2017); Philip Hanke, *Computers with Law Degrees? The Role of Artificial Intelligence in Transnational Dispute Resolution, and Its Implications of the Legal Profession*, 14(2) Transnat'l Disp. Mgmt. 1 (2017).

2 For instance, the part of the US legal service market in relation to new technologies in law is estimated to grow to USD 55 billion (from USD 12 billion in 2017), while at the same time traditional law firm services are estimated to fall to USD 265 billion (from USD 300 billion in 2017). *See* Robert J. Ambrogi et al., *Ethics Issues in Lawyers' Use of Artificial Intelligence*, presentation at 44th ABA National Conference on Professional Responsibility (1 June 2018), www.americanbar.org/content/dam/aba/events/professional_responsibility/2018_cpr_meetings/2018conf/materials/session1_ethics_issues/session1_all_materials.pdf (accessed 9 May 2019).

3 *See* Kate Apostolova & Mike Kung, *Don't Fear AI in IA*, Global Arb. Rev. (27 Apr. 2018); Adesina Temitayo Bello, *Online Dispute Resolution Algorithm: The Artificial Intelligence Model as a Pinnacle*, 84(2) Int'l J. Arb. Mediation & Dispute Mgmt. 159 (2018); Emma Martin, *The Use of Technology in International Arbitration*, in *40 Under 40 International Arbitration* 337–48 (Carlos Gonzalez-Bueno ed., Wolters Kluwer 2018); Paul Cohen & Sophie Nappert, *The March of the Robots*, Global Arb. Rev. (15 Feb. 2017); Sophie Nappert, *Disruption Is the New Black – Practical Thoughts on Keeping International Arbitration on Trend*, (2) ICC Dispute Resolution Bulletin 20, 25–36 (2018); Sophie Nappert, *The Challenge of Artificial Intelligence in Arbitral Decision-Making*, Practical Law UK Articles (4 Oct. 2018); Kathleen Paisley & Edna Sussman, *Artificial Intelligence Challenges and Opportunities for International Arbitration*, 11(1) NYSBA New York Dispute Resolution Lawyer 35 (Spring 2018); Christine Sim, *Will Artificial Intelligence Take over Arbitration?*, 14(1) Asian Int'l Arb. J. 1 (2018); Robert H. Smit, *The Future of Science and Technology in International Arbitration: The Next Thirty Years*, in *The Evolution and Future of International Arbitration* 365–78 (Wolters Kluwer 2016); Francisco Uríbarri Soares, *New Technologies and Arbitration*, VII(1) Indian J. Arb. L. 84 (2018); Gauthier Vannieuwenhuyse, *Arbitration and New Technologies: Mutual Benefits*, 35 J. Int'l Arb. 119–29 (2018); Mohamad S. Abdel Wahab, *Online Arbitration: Traditional Conceptions and Innovative Trends*, in *International Arbitration: The Coming of a New Age?* ICCA Congress Series 17, 654–67 (Albert Jan van den Berg ed., Wolters Kluwer 2013).

4 *See also* Maxi Scherer, *International Arbitration 3.0 – How Artificial Intelligence Will Change Dispute Resolution*, Austrian Y.B. Int'l Arb. 503 (2019). For studies on human arbitral decision-making, *see* in particular Susan D. Franck et al., *Inside the Arbitrator's Mind*, 66 Emory L.J. 1115 (2017).

It will explore whether and how AI can be used to help or potentially replace arbitrators in their task to decide the dispute. Importantly, the subject of this article differs from discussions about online arbitration, which generally refers to proceedings for which processes are streamlined thanks to the use of technology, such as electronic filings, but where human arbitrators remain the decision-makers.[5] Also, while this article focusses on arbitral decision-making, it uses examples and studies from a wide variety of legal areas and its conclusions are relevant for judicial decision-making more globally, not only in international arbitration.

When considering AI for arbitral decision-making, some have speculated about the feasibility of 'robot-arbitrators',[6] but little research has gone into the potential implications of the use of AI in this area. Authors typically either assert that AI is inevitable in the future,[7] or express scepticism, mainly on the assumption that some 'human factor' would be necessary to ensure empathy and emotional justice.[8] This article seeks to explore the topic in a more in-depth fashion, assessing the technical aspects of AI and its implications and limitations, as well as addressing the more fundamental impact it may have on human decision-making processes and theories thereof.

Section 2 defines AI and describes its most important features. A good understanding of the technical aspects of AI is necessary to fully assess its implications for legal decision-making. Section 3 analyses existing studies on the use of AI to predict the outcome of legal decisions. It evaluates their method and results, questioning the extent to which those studies point towards a general applicability of AI for *ex ante* outcome prediction. Section 4 considers the inherent limitations of AI models used, based on the so-called four Vs of Big Data – Volume, Variety, Velocity, and Veracity – and examines their consequences for legal decision-making. In particular, this section discusses the need for sufficient non-confidential case data, the requirement of repetitive fact-patterns and binary outcomes, the problem of policy changes over time, and the risks of bias and data diet vulnerability. Section 5 highlights one

---

5 *See* e.g. Amy J. Schmitz, *Building on OArb Attributes in Pursuit of Justice*, in *Arbitration in the Digital Age* 182 (Maud Piers & Christian Aschauer eds, Cambridge University Press 2018); Pablo Cortés & Tony Cole, *Legislating for an Effective and Legitimate System of Online Consumer Arbitration*, in *Arbitration in the Digital Age*, *supra* n. 5, at 209. For a discussion on the use of arbitration for data disputes, such as those arising out the European General Data Protection Regulation (2016/679) (GDPR), *see* Jacques de Werra, *Using Data Arbitration and Data ADR for Solving Transnational Data Disputes: Lessons from Recent European Regulations?*, Am. Rev. Int'l Arb. (on file with author, forthcoming).

6 Paul Cohen & Sophie Nappert, *Case Study: The Practitioner's Perspective*, in *Arbitration in the Digital Age*, *supra* n. 5, at 126, 140–45. Cohen & Nappert, *supra* n. 3; José María de la Jara, Daniela Palma & Alejandra Infantes, *Machine Arbitrator: Are We Ready?*, Kluwer Arbitration Blog (4 May 2017).

7 Apostolova & Kung, *supra* n. 3.

8 Soares, *supra* n. 3, at 101; de la Jara, Palma & Infantes, *supra* n. 6. *See also* more nuanced Sophie Nappert, *The Challenge of Artificial Intelligence in Arbitral Decision-Making*, Practical Law UK Articles (4 Oct. 2018).

major draw-back of AI decision-making: the difficulty with providing reasoned legal decisions obtained by AI. Section 6 analyses the changes AI decisions would bring for legal theories of judicial decision-making. It shows that AI would change the normative basis for decision-making and thus constitute a significant paradigm shift from a theoretical point of view. Section 7 sets out the conclusions and the main findings of this article.

2 FEATURES OF ARTIFICIAL INTELLIGENCE MODELS

Lawyers often lack basic understanding of artificial intelligence.[9] AI-savvy lawyers are said to be as rare as vegan butchers.[10] Without becoming computer-scientists, it is important for lawyers to understand the basic features of artificial intelligence. Only with a good understanding of AI is it possible to assess its potential implications on the legal profession and legal thinking. The aim of this section is therefore to provide some important technical background information on AI.

Artificial intelligence can be defined as 'making a machine behave in ways that would be called intelligent if a human were so behaving'.[11] This was indeed the definition proposed by John McCarthy, a late computer scientist and arguably the one who coined the term 'AI' in 1956. Other similar definitions exist. For instance, the *Oxford Dictionary* defines artificial intelligence as the '[t]heory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages'.[12]

These definitions show human intelligence as a bench-mark for AI. The term 'intelligence' in itself is not straightforward to define and has caused philosophers, psychologists, cognitive scientists, and other researchers to disagree.[13] At a basic

---

9 Queen Mary School of International Arbitration Survey, *The Evolution of International Arbitration* 33 (2018) ('As far as AI is concerned, the lack of familiarity translates into a fear of allowing technology to interfere excessively with the adjudication function, which is supposed to be "inherently human"').

10 Marc Lauritsen, *Towards a Phenomenology of Machine-Assisted Legal Work*, 1(2) J. Robotics, Artificial Intelligence & L. 67, 79 (2018).

11 John McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (31 Aug. 1955), in *Artificial Intelligence: What Everyone Needs to Know* 1 (Jerry Kaplan ed., Oxford University Press 2016), www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed 9 May 2019).

12 Oxford Living Dictionaries, https://en.oxforddictionaries.com/definition/artificial_intelligence (accessed 9 May 2019).

13 *See* e.g. Shane Legg & Marcus Hutter, *A Collection of Definitions of Intelligence*, 157 Frontiers in Artificial Intelligence & Applications 17 (2007). In the context of AI, the distinction between fluid intelligence (i.e. the ability to reason and think flexibly) and crystallized intelligence (i.e. the accumulation of knowledge, facts, and skills that are acquired throughout life) seems important. *See* e.g. David

level, one can describe intelligence as 'the ability to learn, understand, and make judgments or have opinions that are based on reason'.[14] This ability distinguishes human beings from other forms of non-intelligent or less intelligent life.[15]

In early stages of AI-research, computer scientists tried to develop programs that mimic human intelligence by seeking to understand human cognitive processes and replicate them.[16] For instance, computer scientists tried to understand the processes involved in learning a language and thus develop an algorithm – a sequence of precise instructions – that would enable computers to learn a language. Results were poor, particularly with complex tasks, such as language-learning.[17]

To a lesser extent, similar models are still used today. They are called expert systems or rules-based programs.[18] These systems are based on a set of rules, generally in the form of 'if-then' instructions (e.g. if the light turns red, then stop), also called the knowledge base. They make use of logical inferences, based on the rules contained in the knowledge base. There are several reasons those programs are not as powerful as other models further described below. Most importantly, they are laborious because the knowledge base needs to be created manually by defining the rules and coding the program accordingly.[19] Moreover, the use of *ex ante* rules, such as 'if/then' principles, are often unsuitable to describe accurately complex and dynamic realities.[20]

Different models were thus developed. The quantum leap in AI-research occurred with the so-called 'dataquake', the emergence of huge amounts of data.[21] This surge of data was due to the combination of increased computer processor speed (which is said to double every twelve-eighteen months according to the so-called Moore's Law)[22] and decreased data storage costs (which is said to follow a

---

F. Lohman, *Human Intelligence: An Introduction to Advances in Theory and Research*, 59(4) Rev. Educational Res. 333 (1989).

14 Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/intelligence (accessed 9 May 2019).

15 Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, 24 et seq. (Knopf 2017).

16 Steven Levy, *The AI Revolution Is on*, WIRED (27 Dec. 2010), www.wired.com/2010/12/ff-ai-essay-airevolution (accessed 9 May 2019); Osonde Osoba & William Welser IV, *An Intelligence in Our Image – The Risk of Bias and Errors in Artificial Intelligence* 5 (Rand 2017); Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach* 693 (3d ed., Pearson 2010).

17 Mathias Winther Madsen, *The Limits of Machine Translation* 5–15 (2009) Master Thesis University of Copenhagen, http://vantage-siam.com/upload/casestudies/file/file-139694565.pdf, cited in Harry Surden, *Machine Learning and the Law*, 89 Wash. L. Rev. 87, 99 (2014).

18 Ethem Alpaydin, *Machine Learning* 50–52 (MIT Press 2016); Margaret A. Boden, *Artificial Intelligence: A Very Short Introduction* 26–28 (Oxford University Press 2018).

19 Alpaydin, *supra* n. 18, at 50–52.

20 Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, 55 Communications of the ACM 78, 80 (2012).

21 Alpaydin, *supra* n. 18, at 10–13.

22 Gordon E. Moore, *Cramming More Components onto Integrated Circuits*, Electronics 114 (19 Apr. 1965), reprinted in 86 Proceedings of the Institute of Electrical and Electronics Engineers 82 (1998).

similar pace according to the so-called Kryder's Law).[23] The emergence of 'Big Data' allowed a significant shift in the development of artificial intelligence. Rather than developing complex algorithms for cognitive processes, AI is being used to 'learn' from existing data.

Machine learning refers to a subfield of AI-research concerned with computer programs that learn from experience and improve their performance over time.[24] The reference to 'learning' does not refer to cognitive processes thought to be involved in human learning; rather it refers to the functional sense of learning: the ability to change behaviour through experience over time.[25] The process of machine learning has achieved surprising results in many areas.[26] To continue with the previous example of language-learning, computer translation programs are remarkably accurate these days. Contrary to the earlier attempts described above, no programmer needs to code an algorithm for translation; rather, computer models, such as neural networks, use massive amounts of available data to 'learn' the relevant features and continually improve with immediate online feedback through user clicks. Boden notes that 'many networks have the uncanny property of self-organization from a random start'.[27]

Machine learning at its core relies on the inference of hidden factors or patterns from observed data.[28] Using large amounts of sample data and with sufficient computing power, the computer extracts the necessary algorithms, rather than those algorithms being coded into the machine. In many areas, defining the algorithm in the form of precise *ex ante* instructions proves difficult.[29] For instance, humans might easily recognize which email is spam, but cannot provide precise and exhaustive instructions for this classification task. However, if the program is given a large set of sample data in which emails are labelled as 'spam' or 'not spam', the program will be able to detect the necessary classification algorithm. It does so by recognizing repeat patterns for spam emails and infers that future emails with the same features should also be classified as spam.

The search for hidden patterns is illustrated by the term 'data mining'. The analogy is that one has to work through tons of earth from the mine to find precious material.[30] In the AI context, the program weeds through large amounts of data with the aim to find an accurate model. Once the hidden model is

---

23 Chip Walter, *Kryder's Law*, 293 Scientific American 20 (1 Aug. 2005).
24 Russell & Norvig, *supra* n. 16, at 693.
25 Surden, *supra* n. 17, at 89.
26 For a recent example, *see* a live debate between a human and an AI-driven digital debater, www.research.ibm.com/artificial-intelligence/project-debater/live (accessed 9 May 2019).
27 Boden, *supra* n. 18, at 70.
28 Alpaydin, *supra* n. 18, at xi.
29 Surden, *supra* n. 17, at 94.
30 Alpaydin, *supra* n. 18, at 14.

detected, this can be used to predict future cases (e.g. classify a future email as spam or not), which is of particular importance in the legal context, as further discussed below.[31]

The ability of pattern-recognition relies on statistics and probability calculations.[32] In simple terms, the computer program calculates, for each factor or combination of factors it observes, the probability to lead to a certain outcome. For instance, if the words 'sex' and 'Viagra' are in an email, the probability for it to be spam is high. Probabilistic theories, such as Bayesian networks, are the source of success of machine learning AI.[33] The learning programs resemble a general template with modifiable parameters, with the aim to adapt the parameters of the model on the basis of the information extracted from the sample data. As Alpaydin puts it, in AI '[i]ntelligence seems not to originate from some outlandish formula, but rather from the patient, almost brute force use of simple, straightforward algorithms'.[34]

As a consequence, AI models are able to produce 'intelligent' outcomes which, if performed by humans, are thought to involve high-level cognitive processes (e.g. understanding emails in order to classify them as spam).[35] However, this result is achieved without anything that resembles 'intelligent' human-cognitive processes but is merely based on probabilistic models. As one author describes it, 'research has shown that certain ... tasks can be automated – to some degree – through the use of non-cognitive computational techniques that employ heuristics or proxies (e.g. statistical correlations) to produce useful, "intelligent" results'.[36] The implications for legal decision-making that arise as a result of this shift from early models that focus on human-like processes, to statistical or probabilistic models that achieve human-like results without 'intelligent' processes, is discussed in greater detail below.

AI-researchers distinguish several types of machine learning, depending on the degree of human input. Supervised learning requires human interaction: the programmer trains the program by defining a set of desired outcomes (e.g. classification into spam/no-spam) for a range of input.[37] This means that the data of the training set must be adequately labelled (e.g. emails identified as spam or not)

---

31 *See infra* s. 3.

32 Boden, *supra* n. 18, at 39–40.

33 Alpaydin, *supra* n. 18, at 63–64, 82–84.

34 *Ibid.*, at xii.

35 One early example of 'intelligent' machine behaviour was the IBM 'Deep Blue' computer beating the chess champion Gary Kasparov. On this experiment, which took place already 20 years ago, *see* Gary Kasparov, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins* (John Murray 2017).

36 Surden, *supra* n. 17, at 95.

37 Peter Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data* 2 (Cambridge University Press 2012).

and some form of human feed-back is required (e.g. when the program wrongly classifies an email). To the contrary, unsupervised learning requires no, or virtually no, human interference. There are no pre-established assumptions or pre-defined outputs; rather, the program detects co-occurring features which will engender the expectation that they will co-occur in the future.[38] This is the case, for instance, with many modern language translation programs discussed above.

Importantly, there is not one single AI system, but a variety of different models.[39] For the purpose of the current study, the differences between the two approaches described above are important. On the one hand, expert models are *rule-based* and use *logic* as the normative principle. They may also be described as using a *forward* approach, because they apply pre-established rules to the observable data. The method is *causal*, deducing the outcome from the pre-established, fixed rules coded in the algorithm. On the other hand, machine learning models, such as neural networks, have often no pre-defined rules but use *pattern-recognition* and are built on *probabilistic methods* as the normative principle. They may also be described as using an *inverse* approach, because they extract the algorithm from observable data. The method is *predictive*, calculating the likelihood for any given outcome based on the extracted, and steadily improving, algorithm.

3 LEGAL DECISION-MAKING AND AI: THE USE OF QUANTITATIVE PREDICTION

The idea that AI-driven programs could predict the outcome of legal decision-making seems counter-intuitive to most lawyers. Lawyers instinctively believe that legal decision-making requires cognitive processes – such as understanding the parties' legal submissions and determining the right outcome through reasoning – which cannot be achieved by computer programs. However, as discussed in the previous section, computer models are able to achieve 'intelligent' results, which, if performed by humans, are believed to require high-level cognitive processes.

Several studies may lend support to the thesis that computer programs are better than humans in predicting the outcome of legal decision-making.[40] For instance, an early study showed that computer programs excelled over human

---

[38] Boden, *supra* n. 18, at 40.

[39] For more details, *see ibid.*

[40] For some of the earlier studies, *see* Roger Guimerà & Marta Sales-Pardo, *Justice Blocks and Predictability of U.S. Supreme Court Votes*, 6(11) PloS One (2011); Andrew D. Martin et al., *Competing Approaches to Predicting Supreme Court Decision Making*, 2(4) Persp. Pol. 761 (2004); Theodore W. Ruger et al., *The Supreme Court Forecasting Project: Legal and Political Sciences Approaches to Predicting Supreme Court Decisionmaking*, 104 Colum. L. Rev. 1150 (2004). Generally on forecasting, *see* Philip E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press 2005); Philip E. Tetlock & Dan Gardner, *Superforecasting: The Art and Science of Prediction* (Crown 2015).

experts in predicting the votes of individual US Supreme Court justices in upcoming decisions for the 2002 term. The computer model achieved a correct prediction rate of 75%, whereas the human expert group, composed of eminent lawyers and law professors, correctly guessed only 59.1% of votes.[41]

The basic explanation for this – apparently triumphant – AI-success is that human brains suffer 'hardware' limitations which computer programs surpass easily.[42] In coming years, it is expected that computers available at the consumer level will reach storage capacity of several petabytes. Fifty petabytes are sufficient to store the information content of the 'entire written works of mankind from the beginning of recorded history in all languages'.[43] Accordingly, computers can simply stock amounts of data and draw from that data – or experience – much more quickly and efficiently than humans ever will.[44]

This section discusses two recent studies on the prediction of legal decision-making, looking at their methodology and results. Section 3.1 analyses a study conducted in 2016, which relates to decisions of the European Court of Human Rights, and section 3.2 looks at a study from 2017 predicting US Supreme Court decisions.

3.1 Predicting decisions of the European Court of Human Rights

The study conducted by a group of researchers in 2016[45] focussed on decisions by the European Court of Human Rights (hereafter the 'ECtHR') rendered in the English language about three provisions of the European Convention on Human Rights (hereafter the 'Convention'),[46] namely Article 3 on the prohibition of torture, Article 6 on the right to a fair trial, and Article 8 on the right to respect for private and family life. Those provisions were chosen because they provided the highest number of decisions under the Convention and thus sufficient data on

41 Ruger et al., *supra* n. 40, at 1152.

42 Tegmark, *supra* n. 15, at 27–28.

43 *How Much Is a Petabyte?*, Mozy BLOG (2009), cited in Daniel M. Katz, *Quantitative Legal Prediction*, 62 Emory L.J. 909, 917 (2013).

44 Interestingly, France has recently prohibited, under threat of criminal sanctions, the use of certain data from published decisions for predictive analytics. A newly introduced provision states that '[t]he identity data of magistrates and members of the judiciary cannot be used with the purpose or effect of evaluating, analysing, comparing or predicting their actual or alleged professional practices'. *See* Law No. 2019-222 (23 Mar. 2019), Art. 11, www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000038262498&dateTexte=20190604 (accessed 9 May 2019).

45 Nikolaos Aletras et al., *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective*, PeerJ Computer Science 2:e93 (2016).

46 The 'European Convention on Human Rights' refers to the Convention for the Protection of Human Rights and Fundamental Freedoms, signed in Rome on 4 November 1950, as amended and supplemented by subsequent Protocols Nos. 1, 4, 6, 7, 12, 13, 14, and 16, www.echr.coe.int/Documents/Convention_ENG.pdf (accessed 9 May 2019).

which to base a study.[47] For each of those provisions, the study selected an equal number of decisions in which the ECtHR found a violation and in which it found none. This resulted in a total dataset of 584 decisions: 250 for Article 3, 80 for Article 6, and 254 for Article 8.[48]

The methodology used in the study focussed on the textual information contained in the decisions, using natural language processing and machine learning.[49] The study input was the text found in the decisions, following the usual structure of decisions of the ECtHR including sections on the procedure, factual background, and legal arguments.[50] Not included in the input were the operative sections of the decisions where the Court announces the outcome of the case.[51] The output target was a binary classification task as to whether or not the ECtHR found a violation of the underlying provision of the Convention.[52] The model was trained and tested on a 10% subset of the dataset.[53]

As a result, the model obtained an overall accuracy to predict the outcome of the Court's decision in 79% of all cases.[54] The decision sections with the best predictive value were those setting out the factual circumstances and procedural background (76% and 73%, respectively), whereas the legal reasoning section had a lesser outcome prediction value (62%).[55] The study also set out the most frequently used words for various topics, indicating their relative predictive weight for a violation or non-violation. For instance, the most frequently used words with a high prediction value included under Article 3 of the Convention: 'injury', 'damage', 'Ukraine', 'course', 'region', 'effective', 'prison', 'well', 'ill treatment', 'force', and 'beaten'[56]; under Article 6 of the Convention: 'appeal', 'execution', 'limit', 'copy', 'employee', 'January', and 'fine'[57]; and under Article 8 of the Convention: 'son', 'body', 'result', 'Russian', 'department', 'attack', and 'died'.[58]

The authors of the study claim that their work may lead the way to predicting *ex ante* the outcome of future ECtHR cases. They state that:

> [o]ur work lends some initial plausibility to a text-based approach with regard to *ex ante* prediction of ECtHR outcomes on the assumption that the text extracted from published judgments of the Court bears a sufficient number of similarities with, and can therefore

---

47 Aletras et al., *supra* n. 45, at 6.
48 *Ibid.*, at 8.
49 *Ibid.*, at 1.
50 *Ibid.*, at 4–6.
51 *Ibid.*, at 8.
52 *Ibid.*, at 2.
53 *Ibid.*, at 9.
54 *Ibid.*, at 10.
55 *Ibid.* In addition to the decision sections, the study also created certain topics, which overall had a higher prediction result than the decision sections and which, combined, led to the overall result of 79%.
56 *Ibid.*, at 13, table 3.
57 *Ibid.*, at 14, table 4.
58 *Ibid.*, at 15, table 5.

> stand as a (crude) proxy for, applications lodged with the Court as well as for briefs submitted by parties in pending cases.[59]

The authors further see in the above-mentioned results a confirmation of legal realist theories according to which judges are primarily responsive to non-legal, rather than to legal, reasons when deciding cases.[60] They conclude that 'the information regarding the factual background of the case as this is formulated by the Court in the relevant subsections of its judgment is the most important part obtaining on average the strongest predictive performance of the Court's decision outcome' and thus suggest that 'the rather robust correlation between the outcomes of cases and the text corresponding to fact patterns ... coheres well with other empirical work on judicial decision-making in hard cases and backs basic legal realist intuitions'.[61] The conclusion on the validation of legal realists' theories will be discussed in detail in section 7 below. This section provides some comments on the methodology and results, as well as the claim that the study leads the way to *ex ante* outcome prediction.

First, it remains somewhat unclear which parts of the ECtHR decisions were included in the study's input. As indicated above, the operative part of the decision in which the Court announces the outcome of the case, is obviously not included,[62] otherwise the prediction-task would be moot. Less clear is whether the part of the legal section containing the Court's reasoning is included or not. The study indicates that the aim was to 'ensure that the models do not use information pertaining to the outcome of the case' but this caveat seems to apply only to the operative sections of the decisions.[63] The law section is said to be included[64] and this typically includes the Court's legal reasoning, as indicated in the study.[65]

If the Court's legal reasoning is indeed included in the data input, the study's overall prediction results are all but surprising. Any trained lawyer – and probably most non-lawyers – would be able to guess, in virtually 100% of the cases, the outcome as to whether the Court finds a violation or not, after having been given the Court's reasoning. The study's overall prediction rate of 79% is therefore to be interpreted in this context. Moreover, the inclusion of the Court's legal reasoning significantly undermines the study's claim to lead the way towards possible *ex ante* outcome prediction. The Court's reasoning is precisely not available *ex ante* and therefore cannot be included in the prediction of future cases.

---

59 *Ibid.*, at 2.
60 *Ibid.*, at 12.
61 *Ibid.*, at 16.
62 *Ibid.*, at 8.
63 *Ibid.*
64 *Ibid.*, at 8, 10.
65 *Ibid.*, at 5.

Second, one may query whether the factual background part in the Court's decision does not already contain 'hints' concerning the decision's outcome. The study acknowledges the 'possibility that the formulation by the Court may be tailor-made to fit a specific preferred outcome'.[66] Without suggesting any form of bias or lack of neutrality on the part of the ECtHR judges, the facts described in the judgment may be a selection of those facts that will be relevant for the decision's legal reasoning and outcome, leaving aside other non-pertinent facts pleaded by the parties. Therefore, one may express doubts as to the study's assumption that 'the text extracted from published judgments of the Court bears a sufficient number of similarities with, and can therefore stand as a (crude) proxy for, applications lodged with the Court as well as for briefs submitted by parties in pending cases'.[67]

Third, the most frequently used words for various topics with a high prediction value set out in the study would have to be used in any *ex ante* prediction model. This seems problematic for a number of reasons. Some of the words – such as 'result', 'employee', 'region', 'copy', or 'department' – seem random and it is hard to see how they would be able to predict *ex ante* the outcome of future cases. Others are very case-specific and would be problematic if used for future predictions, including words such as 'Ukraine', 'January', or 'Russian'. Using these words for future outcome prediction might lead to facts relating to those countries or dates being determinative on the outcome. Implications of possible text-based prediction tools are further discussed below.[68]

Overall, while the result of the study, obtaining 79% accuracy to predict the outcome of the ECtHR decisions, seems impressive at first sight, a closer analysis of the methodology and assumptions employed puts into doubt the claims for possible *ex ante* outcome predictions.

3.2 Predicting decisions of the US Supreme Court

Another group of researchers focussed on the prediction of US Supreme Court decisions and published their final results in 2017.[69] Their study drew from previous work on US Supreme Court predictions,[70] but was innovative in several

---

66 *Ibid.*

67 *Ibid.*, at 2.

68 *See infra* ss 4.2 and 5.

69 Daniel M. Katz, Michael J. Bommarito II & Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, 12(4) PloS One (2017).

70 Guimerà & Sales-Pardo, *supra* n. 40; Martin et al., *supra* n. 40; Ruger et al., *supra* n. 40. *See also* Michael A. Bailey & Forrest Maltzman, *Does Legal Doctrine Matter? Unpacking Law and Policy Preferences on the U.S. Supreme Court*, 102(3) Am. Pol. Sci. Rev. 369 (2008); Stuart M. Benjamin & Bruce A. Desmarais, *Standing the Test of Time: The Breadth of Majority Coalitions and the Fate of U.S. Supreme Court*

aspects. First, the study's goal was to obtain a model that would generally and consistently be applicable to all US Supreme Court decisions over time, not only in a given year or for a given composition of the Court with justices.[71] Second, the study also applied the principle that 'all information required for the model to produce an estimate should be knowable prior to the date of the decision'.[72] As has been discussed in the previous section, this is to ensure that the model can be used for *ex ante* outcome prediction.

In order to achieve these aims, the study input included US Supreme Court decisions from almost two centuries, from 1816 to 2015. This resulted in input data of more than 28,000 case outcomes and more than 240,000 individual justices' votes.[73] Rather than relying on the textual information contained in the decisions themselves, as was the case for the ECtHR study, this study labelled the data relating to each decision, using certain features.[74] First, some features relate to the specific case at hand, such as the identity of the parties, the issues at stake or the timing of the decision to be rendered. Second, other features draw information from the lower court's decision which is to be examined. This includes, among others, the identity of the courts of origin (i.e. which circuit), the lower court's disposition and directions, as well as which lower courts are in disagreement over the issue at stake. Third, another category of features focusses on the Supreme Court's composition, such as the identity of the justices, and their previous rate of reversal votes or dissents, as well as their political preferences. Fourth, a final set of features relates to the procedure before the US Supreme Court, such as the manner in which the Court took jurisdiction and the reasons for granting certiorari,[75] whether or not an oral argument was scheduled and, if so, the time between the argument and the decision.

Precedents, 4 J. Leg. Analysis 445 (2012); Lee Epstein et al., *Ideological Drift Among Supreme Court Justices: Who, When, and How Important*, 101 Nw. U. L. Rev. 1483 (2007); Edward D. Lee, Chase P. Broedersz & William Bialek, *Statistical Mechanics of the US Supreme Court*, 160 J. Statistical Physics 275 (2015); Andrew D. Martin & Kevin M. Quinn, *Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999*, 10(2) Pol. Analysis 134 (2002); Jeffrey A. Segal & Harold J. Spaeth, *The Supreme Court and the Attitudinal Model Revisited* (Cambridge University Press 2002); Jeffrey A. Segal & Harold J. Spaeth, *The Influence of Stare Decisis on the Votes of United States Supreme Court Justices*, 40 Am. J. Pol. Sci. 971 (1996); Jeffrey A. Segal et al., *Ideological Values and the Votes of U.S. Supreme Court Justices Revisited*, 57(3) J. Pols. 812 (1995); Carolyn Shapiro, *Coding Complexity: Bringing Law to the Empirical Analysis of the Supreme Court*, 60 Hastings L.J. 477 (2008).

71 Katz, Bommarito & Blackman, *supra* n. 69, at 2–3.

72 *Ibid.*, at 3.

73 *Ibid.*, at 5, table 1.

74 For a full list of the features, *see ibid.*, at 4–6. Many of the features used were not originally labelled but taken from the Supreme Court Database (SCDB) established by Harold Spaeth for the use of empirical studies. Harold J. Spaeth et al., *Supreme Court Database* (Version 2016, Legacy Release v01 (SCDB Legacy 01)), supremecourtdatabase.org (accessed 9 May 2019).

75 A petition for a writ of certiorari is the most common procedural device to invoke the US Supreme Court's appellate jurisdiction. *See* 28 U.S.C. § 1254(1), § 1257, § 1259. *See also* Steven M. Shapiro et al., *Supreme Court Practice* 59 et seq. (10th ed. 2013).

The study output target was two-fold: predicting the outcome of the decisions and predicting each justice's votes.[76] For the outcome of the decisions, the classification task was binary, as to whether the Supreme Court reversed or affirmed the lower court's decision.[77] There are some (albeit few) cases in which the Supreme Court does not review a lower court's decision, but rather decides a dispute as the original court of jurisdiction.[78] The study excluded those cases from the decision outcome prediction because they do not fall into a binary classification task.[79]

Using machine learning, the researchers trained the model on a sample from the dataset, and then applied the obtained model to the remaining, out-of sample, data.[80] Overall, the model predicted the votes of individual justices with 71.9% accuracy, and the outcome of the decisions with 70.2% accuracy.[81] While there was fluctuation in any given year or decade, the study claims that the model delivered 'stable performance' over time.[82] The study also claims that the model 'significantly outperforms' possible baseline comparison models.[83]

Testing the study's methodology and results against its aim to provide a general model for *ex ante* outcome prediction, the study contains some important limitations.

First, while the study applies the principle that 'all information required for the model to produce an estimate should be knowable prior to the date of the decision',[84] some of the input data features are available only shortly before the decision is rendered. For instance, whether or not an oral argument is scheduled and, if so, the time between the argument and the decision, is information typically available only at a late stage of the proceedings.[85] This significantly limits the use of those features for *ex ante* outcome prediction.

Second, a majority of the input-data labels are specific to appellate or Supreme courts tasked with the review of lower courts' decisions. As detailed above, many features used in the study are related to the lower court's decision to be examined

---

76 Katz, Bommarito & Blackman, *supra* n. 69, at 4.

77 *Ibid.*

78 The US Supreme Court has original (i.e. acts as a court of first instance) exclusive jurisdiction over controversies between States, and concurrent original jurisdiction over proceedings involving ambassadors and certain other foreign officials, controversies between the United States and a State, and proceedings by a State against citizens of another State or aliens. 28 U.S.C. § 1251; *see also* US Const., Art. III, § 2.

79 Katz, Bommarito & Blackman, *supra* n. 69, at 4.

80 *Ibid.*, at 6–8.

81 *Ibid.*, at 8–9.

82 *Ibid.*, at 9.

83 *Ibid.*, at 15.

84 *Ibid.*, at 3.

85 The study notes that 'in practice, the predictions for a case may evolve as new information about the case is acquired prior to the decision being rendered'. *Ibid.*, at 5.

(e.g. which circuit, the lower court's disposition and directions) as well as the Supreme Court justice's handling of previous decisions from lower courts (e.g. reversal rates). Few of the input features are original to the dispute, such as the identity of the parties, the issues at stake or procedural aspects before the decision is rendered. Accordingly, it is questionable whether the methodology or model may equally apply and provide successful results for cases where the court originally decides a dispute, rather than reviewing a lower court's decision.

Third, and somewhat relatedly, the decision outcome prediction only applies for the binary classification tasks as to whether the Supreme Court reverses or affirms the lower court's decision. As mentioned above, cases in which the Supreme Court decides a dispute as the original court of jurisdiction are excluded from the study. The study notes that this is so because 'the Court and its members may take technically nuanced positions or the Court's decision might otherwise result in a complex outcome that does not map onto a binary outcome'.[86] The very same may be said about most instances in which a court originally decides a dispute, rather than reviewing another court's decision. In those cases, the court will have to decide technically complex and nuanced matters of facts and law which are difficult to classify into a binary model. The issue of binary-tasks for AI models are further discussed below.[87] At this stage, suffice it to note that the study's methodology is not easily transposable to lower courts' decisions whose task is to originally decide a dispute rather than reviewing another court's previous decisions.

Fourth, one might also note that decisions of Supreme courts generally, and of the US Supreme Court in particular, are often highly political. US Supreme Court's justices are indeed appointed considering their political orientation, among other things.[88] The points of law on which the US Supreme Court renders decisions are often those on which lawyers from different sides of the political spectrum come out differently, the possibility of gun control being one example.[89] To the contrary, lower courts' decision are typically more fact-driven and less legally principled. Some of the features used (e.g. the judge's political orientation) are therefore less likely to be outcome-determinative, or at least the relation between the feature and the outcome is not going to be as straight-forward.

---

86 *Ibid.*

87 *See infra* s. 4.

88 Neal Devins & Lawrence Baum, *Split Definitive: How Party Polarization Turned the Supreme Court into a Partisan Court*, 2016 Sup. Ct. Rev. 301, 331 (2016).

89 *See District of Columbia v. Heller*, 554 U.S. 570 (2008); *McDonald v. Chicago*, 561 U.S. 742 (2010). However, challenging the assumption that US Supreme Court justices vote on the basis of one-dimensional policy preference, *see* Joshua Fischman, *Do the Justices Vote Like Policy Makers? Evidence from Scaling the Supreme Court with Interest Groups*, 44 J. Legal Stud. S269 (2015).

Overall, the above-mentioned studies therefore have important inherent limitations as to their general applicability for *ex ante* outcome prediction. Nevertheless, they spark the questions as to whether AI-driven and machine learning based outcome prediction tools might not be a useful addition to human decision-making. Max Radin wrote in 1925 about judicial decision-making that the judge's 'business is prophecy, and if prophecy were certain, there would not be much credit in prophesying'.[90] If AI models could prophesize or help with predictions, should they not replace, or at a minimum be taken into account by, human decision-makers? The following sections of this article aim at helping to provide an answer to this question.

4 LIMITATIONS ON LEGAL DECISION-MAKING WITH AI: THE FOUR 'V'S OF BIG DATA

Data specialists often refer to the four Vs of Big Data – Volume, Variety, Velocity, and Veracity – as the cornerstones of data-driven projects.[91] The four Vs describe challenges to Big Data use. They also help to assess data-driven AI programs such as those described in the previous section, and their use in the legal sector. This section looks at the four Vs in turn and discusses inherent limitations of data-driven models for legal decision-making with AI.

4.1 VOLUME: NEED FOR SUFFICIENT NON-CONFIDENTIAL CASE DATA

Any data-driven AI programs first and foremost require access to data. Machine learning models, which are based on probabilistic inferences, are data-hungry: the larger the sample data, the more accurate the model's predictive value. In the legal sector, the volume of data required leads to a possible two-fold limitation of AI programs.

First, case data is not always easily accessible. In certain areas of law, decisions are confidential and thus not available to non-parties. Confidentially can be based on protecting the affected parties' rights or the underlying transactions. For instance, international commercial arbitration awards are generally not published and the constitution of a database to establish an AI model would therefore prove

---

90 Max Radin, *The Theory of Judicial Decision: Or How Judges Think*, 11 ABA J. 357, 362 (1925).

91 Initially, the focus was on only three Vs (volume, variety, and velocity). *See* e.g. Max N. Helveston, *Consumer Protection in the Age of Big Data*, 93 Wash. U. L. Rev. 859, 867 (2016). Veracity was added in the mid-2000s. *See also* Margaret Hu, *Small Data Surveillance v. Big Data Cybersurveillance*, 42 Pepp. L. Rev. 773, 795 (2015); Todd Vare & Michael Mattioli, *Big Business, Big Government and Big Legal Questions*, 243 Managing Intell. Prop. 46 (2014). More recently, some have suggested a fifth V in the form of 'value'. *See* e.g. Amy Affelt, *Big Data, Big Opportunity*, 21 Austl. L. Libr. 78 (2013). In the legal context, this last point is of less relevance and thus not discussed here.

difficult.[92] However, this is not to say that AI models in international commercial arbitration are impossible. Initiatives exist to publish commercial awards on a regular basis, typically in a redacted format.[93] In any event, even without publishing confidential awards, institutions could collect them and make them available for the purpose of building AI models.

Second, when case data is accessible, a large sample size is important. While there is no hard rule of a required sample size, the more data, the more accurate the extracted model. Accordingly, areas of law with large numbers of decisions on a given topic will be more suitable for AI models. In international investment arbitration, although there are no reliable statistics on how many awards are rendered per year, on the basis that around sixty new cases are initiated per year,[94] the number of arbitral awards should similarly only be in the double-digits,[95] which does not make for a particularly sample size.

4.2 Variety: Requirement of repetitive patterns with binary outcomes

In addition to the necessary data volume, there is also a question about the variety of the input data. In data-research terminology, variety of data refers to the fact that data comes from different sources and may be structured (e.g. a file containing names, phone numbers, addresses) or unstructured (photos, videos, social media feeds).[96] In the legal context, the variety question is likely to be framed in a different manner. The variety will not so much come from different sources or formats – since the input data is likely to be limited to previous decisions – but

---

92 Queen Mary School of International Arbitration Survey, *The Evolution of International Arbitration* 3, 24 (2018) ('87% of respondents believe that confidentiality in international commercial arbitration is of importance'); Queen Mary School of International Arbitration Survey, *Improvements and Innovations in International Arbitration* 6 (2015) (respondents cited 'confidentiality and privacy' as one of the top five most valuable characteristics of international arbitration, with the in-house counsel subgroup rating it as the second most valuable characteristic).

93 *See* e.g. ICC, *Note to Parties and Arbitral Tribunals on the Conduct of the Arbitration Under the ICC Rules of Arbitration*, paras 42–43 (1 Jan. 2019), https://cdn.iccwbo.org/content/uploads/sites/3/2017/03/icc-note-to-parties-and-arbitral-tribunals-on-the-conduct-of-arbitration.pdf (accessed 9 May 2019).

94 According to UNCTAD statistics, sixty-two new treaty-based investor–State dispute settlement cases were initiated in 2016, sixty-five in 2017 and at least seventy-one in 2018. *See* UNCTAD, *Investor-State Dispute Settlement: Review of Developments in 2016* 1 (May 2017); UNCTAD, *Investor-State Dispute Settlement: Review of Developments in 2017* 1 (June 2018); UNCTAD, *New ISDS Numbers: Takeaways on Last Year's 71 Known Treaty-Based Cases* (13 Mar. 2019), https://investmentpolicyhubold.unctad.org/News/Hub/Home/1609 (accessed 9 May 2019).

95 This takes into account that, on the one hand, some disputes will settle without any award being rendered, and on the other hand, some disputes will rise to multiple partial awards.

96 *See* e.g. EY, *Big Data: Changing the Way Businesses Compete and Operate*, Rpt. 2 (Apr. 2014); Lieke Jetten & Stephen Sharon, *Selected Issues Concerning the Ethical Use of Big Data Health Analytics* 72 Wash. & Lee L. Rev. Online 486, 487 (2016); Uthayasankar Sivarajah et al., *Critical Analysis of Big Data Challenges and Analytical Methods*, 70 J. Bus. Research 263, 269 (2017).

rather from the content dealt with in those decisions. For AI-driven decision-making two variety questions come to mind.

The first question relates to the data input and to what extent AI-based decision-making models require repetitive fact patterns or, conversely, whether they would be able to deal with topics that are complex and non-repetitive. In the above-mentioned study on US Supreme Court decisions, the computer program was developed for decisions spanning over almost two hundred years and dealing with a large variety of issues.[97] Nevertheless, the more outliers or non-repetitive issues, the more difficulties the AI model will face. In international arbitration, therefore, AI programs are more likely to apply to international investment arbitration (which typically raises a number of well-known issues) than in international commercial arbitration (which deals with diverse and often unique issues).

The second question relates to the model output. The legal prediction studies discussed above all use a binary classification as the output task. In the case of the ECtHR, the binary classification was whether or not a violation of the relevant provision of the Convention occurred, and in case of the US Supreme Court decision, the binary classification task was whether or not the Court affirmed the lower court's decision. As already noted above, this raises the question whether those, or other similar models, could be built for more diverse, non-binary tasks.[98]

One might be tempted to reply that any legal decision could be subdivided into a multitude of binary classification tasks, such as whether (1) the tribunal has jurisdiction: yes/no; (2) the parties validity entered into a contract: yes/no; (3) one party breached the contract: yes/no etc. Lord Hoffman has famously described a standard of proof issue using a binary analogy:

> If a legal rule requires a fact to be proved (a 'fact in issue'), a judge or jury must decide whether or not it happened. There is no room for a finding that it might have happened. The law operates a binary system in which the only values are 0 and 1. The fact either happened or it did not. If the tribunal is left in doubt, the doubt is resolved by a rule that one party or the other carries the burden of proof. If the party who bears the burden of proof fails to discharge it, a value of 0 is returned and the fact is treated as not having happened. If he does discharge it, a value of 1 is returned and the fact is treated as having happened.[99]

However, while it is true that many legal questions of fact or law can be reduced to a 0/1 or yes/no binary task, the problem is that there will be a multitude of such binary tasks in each case, and determining all of them will be case-specific. For an AI model to be able to extract the required patterns and algorithms from the input data, having one clear output question facilitates the model-building process. This

---

97 *See supra* s. 3.2.
98 *See supra* s. 3.2.
99 *In re B* [2008] UKHL 35.

is why, in the study on US Supreme Court decisions, the research group specifically excluded those decisions in which the Supreme Court was the court of original jurisdiction, which did not correspond to a simple binary classification task.[100]

4.3 Velocity: Problem of policy changes over time

Velocity refers to the frequency of incoming data that needs to be processed. Big Data is often challenging because of the sheer amount and high frequency of the incoming data. In the legal context, such risk is very low. As already pointed out above, in terms of volume, the problem is likely to be of scarcity rather than abundance of data.[101] Therefore, over time, decisions might not be frequent, and when they occur there might have been a change in policy so that the previous data is outdated. These policy changes can be radical and swift at times. To take an example from the international arbitration context, the decision of the Court of Justice of the European Union in *Achmea* has fundamentally changed the compatibility of investor-state arbitration with European law overnight.[102]

This raises the question how AI models which, by definition, are based on information extracted from previous data may deal with those policy changes. It is true that the essence of machine learning is the ability to improve the algorithm over time. Nevertheless, such improvement is always based on past data. Policy changes in case law necessarily require departures from past data, i.e. previous cases. For these reasons, AI models are likely to keep 'conservative' approaches that are in line with previous cases.

4.4 Veracity: Risk of bias and data diet vulnerability

Finally, veracity relates to the accuracy and trustworthiness of the data used. In the AI context, the question is whether there are any hidden data vulnerabilities which might affect the model's accuracy. The robustness and trustworthiness of AI are recurrent topics in the discussion on AI.[103]

As a starting point, one might assume that AI models have the advantage of algorithmic objectivity and infallibility over humans who inevitably make mistakes and are influenced by subjective, non-rational factors. Research in the

---

[100] *See supra* s. 3.2.
[101] *See supra* s. 4.1.
[102] Case C-284/16 *Slovak Republic v. Achmea B.V.* (CJEU, 6 Mar. 2018).
[103] *See* e.g. European Commission Press Release, *Artificial Intelligence: Commission Takes Forward Its Work on Ethics Guidelines* (8 Apr. 2019), http://europa.eu/rapid/press-release_IP-19-1893_en.htm (accessed 9 May 2019).

area of psychology, cognitive science, and economy has shown that humans often fail to act rationally.[104] Most famously, Nobel-prize winner Daniel Kahneman and Amos Tversky have studied heuristics and cognitive biases in human choices.[105] Their studies provide multiple examples in which heuristics (i.e. cognitive short-cuts for otherwise intractable problems) and biases (i.e. factors which appear to be irrelevant to the merit of our choices but affect them nonetheless) appear in human day-to-day decisions.[106]

Applying this research in the legal sector, a group of Israeli and US researchers have shed some light on the importance of extraneous factors in judicial decision-making.[107] Looking at more than 1,100 decisions rendered over ten months by Israeli judges in relation to 40% of the country's parole applications,[108] the study showed that the majority of applications are rejected on average,[109] but the probability of a favourable decision is significantly higher directly after the judge's daily food breaks.[110] While not falling into the generalization of the well-known saying that 'justice is what the judge had for breakfast', the results 'suggest that judicial decisions can be influenced by whether the judge took a break to eat'.[111] This research provides an empirical example about how human decision-making is affected by extraneous factors, such as food breaks, which ought to be irrelevant to the merit of the case.[112]

---

104 *See* e.g. Christine Jolls, Cass Sunstein & Richard Thaler, *A Behavioral Approach to Law and Economics*, 50 Stan. L. Rev. 1471 (1998); Avishalom Tor, *The Methodology of the Behavioral Analysis of Law*, 4 Haifa L. Rev. 237 (2008). Regarding the idea of ecological rationality (rationality is variable and depends on the context), *see* e.g. Vernon L. Smith, *Constructivist and Ecological Rationality in Economics*, 93(3) Am. Econ. Rev. 456 (2003).

105 *See* e.g. Daniel Kahneman & Amos Tversky, *Subjective Probability: A Judgment of Representativeness*, 3 Cognitive Psychol. 430, 431 (1972); Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 Science 1124 (1974); Amos Tversky & Daniel Kahneman, *Availability: A Heuristic for Judging Frequency and Probability*, 5 Cognitive Psychol. 207 (1973). Further research has emphasized the fact that the use of intuitive, non-rational decision-making is both a source of error and a factor of success for humans in their daily choices, and that humans have at least an intuitive logical and probabilistic knowledge. *See* e.g. Wim De Neys, *Bias and Conflict: A Case for Logical Intuitions*, 7(1) Persps Psychological Sci. 28 (2012); Jonathan Evans & Keith E. Stanovich, *Dual-Process Theories of Higher Cognition Advancing the Debate*, 8(3) Persps Psychological Sci. 223 (2013).

106 For instance, a series of studies on the so-called anchor-effect has shown that people, when estimating an unknown quantity, are affected by a number given to them, even if it is obvious that this number is random. *See* Daniel Kahneman, *Thinking, Fast and Slow* 119–128 (Penguin 2011). *See also* Edna Sussman, *Biases and Heuristics in Arbitrator Decision-Making: Reflections on How to Counteract or Play to Them*, in *The Roles of Psychology in International Arbitration* (Tony Cole ed., Wolters Kluwer 2017).

107 Shai Danziger et al., *Extraneous Factors in Judicial Decisions*, 108(17) PNAS 6889 (2011).

108 Parole is a permanent release of a prisoner who agrees to certain conditions before the completion of the maximum sentence period.

109 *Ibid.*, at 6889 (64.2% of the applications in the sample were rejected).

110 *Ibid.*, at 6890 (the probability of parole being granted spikes at approximately 0.65 at the beginning of the session after each food break and declines to nearly 0 at the end of each session).

111 *Ibid.* More specifically, the study concludes that judges when making repeat rulings show a tendency to rule in favour of the status quo (i.e. reject the parole application for liberation) and that this tendency can be overcome, for instance, by taking a food break. *Ibid.*, at 6892.

112 *See also* Chris Guthrie, Jeffrey Rachlinski & Andrew J. Wistrich, *Inside the Judicial Mind*, 86 Cornell L. Rev. 777 (2001).

Some authors have therefore concluded that AI-based decision-making would be superior to human decision-making on the basis that computers would be immune to cognitive biases or undue influence of extraneous factors.[113] However, a blind deferential attitude towards algorithmic objectivity and infallibility is misplaced. AI research over the past years has highlighted the risks of misbehaving or biased algorithms. Important studies discuss bias concerns in computer systems used for a variety of tasks, such as flight listings, credit scores, or on-line advertisements.[114] Referring to a 'scored society', some have argued that hidden and unregulated algorithms produce authoritative scores of individuals that mediate access to opportunities.[115] As other authors put it, 'procedural consistency is not equivalent to objectivity'.[116]

Any data-based computer models are only as good as the input data. Vulnerability in the data diet has negative consequences on the extracted model. In particular, the underlying data which was used to train the algorithm might have been 'infected' with human biases. The machine learning algorithm will be based on those biases and possibly even exaggerate them by holding them as 'true' for its future decisions or outcome predictions.

For instance, in the area of investment arbitration, concerns have been voiced that arbitral tribunals are inherently and unduly investor-friendly.[117] I do not discuss here whether this criticism is well-founded,[118] but rather assume for the purpose of the present demonstration that such human bias exists. In this case, an AI model based on investment arbitration data would be likely to perpetuate such (alleged)

---

[113] Hanke, *supra* n. 1, at 8.

[114] *See* e.g. Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM Transactions on Information Systems 330 (1996); Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms* (paper presented to the Data and Discrimination: Converting Critical Concerns into Productive Inquiry Preconference of the 64th Annual Meeting of the International Communication Association, 22 May 2014); Latanya Sweeney, *Discrimination in Online Ad Delivery*, 11(3) ACM Queue 10 (2013); Nicholas Diakopoulos, *Algorithmic Defamation: The Case of the Shameless Autocomplete*, Nick Diakopoulos (6 Aug. 2013), www.nickdiakopoulos.com/2013/08/06/algorithmic-defamation-the-case-of-the-shameless-autocomplete (accessed 9 May 2019).

[115] Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1 (2014).

[116] Osoba & Welser IV, *supra* n. 16, at 2.

[117] *See* e.g. Pia Eberhardt et al., *Profiting from Injustice: How Law Firms, Arbitrators and Financiers Are Fuelling an Investment Arbitration Boom* 8 (Corporate Europe Observatory 2012); George Kahale III, *Is Investor-State Arbitration Broken?*, 9(7) Transnat'l Disp. Mgmt. 1, 1–2 (2012); Gus van Harten, *Part IV Chapter 18: Perceived Bias in Investment Treaty Arbitration*, in *The Backlash Against Investment Arbitration* 433 (Michael Waibel et al. eds, Wolters Kluwer 2010).

[118] *See* e.g. Gloria Maria Alvarez et al., *A Response to the Criticism Against ISDS by EFILA*, 33(1) J. Int'l Arb. 1, 4 (2016); Carolyn B. Lamm & Karthik Nagarajan, *The Continuing Evolution of Investor-State Arbitration as a Dynamic and Resilient Form of Dispute Settlement*, V(2) Indian J. Arb. L. 93, 96–97 (2016); Stephen M. Schwebel, *Keynote Address: In Defence of Bilateral Investment Treaties*, in *Legitimacy: Myths, Realities, Challenges*, 18 ICCA Congress Series 1, 6 (Albert Jan van den Berg ed., Wolters Kluwer 2015).

favour given to investors. The model would likely predict favourable outcomes for investors against States in a disproportionate number of cases.

Even without going as far as pointing towards human biases in the underlying data, the model might extract patterns from the data and extrapolate them in a way that might lead to systemic mistakes. For instance, studies have shown that the use of algorithms in criminal risk assessment in the United States has led to racially biased outcomes.[119] The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system is widely used in the United States to assess the recidivism risks for defendants. Under this system, studies found that '[b]lack defendants were ... twice as likely as white defendants to be misclassified as a higher risk of violent recidivism', whereas 'white violent recidivists were 63% more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists'.[120] Whether this racial bias in the computer program was based on existing human biases in the training data remains unclear. It might also have resulted from the fact that the algorithm wrongly classified black defendants at the higher recidivist rate because this racial group is overrepresented in certain kinds of crimes. The computer model might have extrapolated from this pattern the wrong assumption of a higher recidivist risk.

The occurrence of systemic errors based on hidden patterns in the underlying data is a serious risk. As discussed above, in the study on ECtHR decisions, words with high predictive value include 'Ukraine' or 'Russian'.[121] Presumably, this was the case because a significant number of ECtHR cases are directed and decided against these countries.[122] Statistics show that a number of countries receive the most applications and condemnations.[123] A computer program modelled on data containing a higher proportion of condemnations of a given country might extrapolate a higher risk of a violation committed by this country in the future and its outcome predictions might thus be biased against this country.

---

[119] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, ProPublica (23 May 2016), www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 9 May 2019); Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, ProPublica (23 May 2016), www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed 9 May 2019).

[120] Jeff Larson et al., *supra* n. 19, at 2.

[121] *See supra* s. 3.1.

[122] In the above-mentioned study on ECtHR decisions, the research group specifically selected the same number of violation and non-violation decisions in order not to pre-influence the data model in one way or another. However, no such precaution was taken when it comes to other criteria, such as geographical origin of parties. *See supra* s. 3.1.

[123] European Court of Human Rights, *Violations by Article and by State, 1959–2018* (2018) (finding that Turkey and the Russian Federation lead the list of countries with most judgments having found at least one violation of the Convention).

It is therefore important to consider whether and how systemic mistakes in algorithms might be addressed. In systems where the algorithm is coded by a human programmer, the mistake will often be in the design of the algorithms itself. It can be changed once the mistake is detected.[124] To the contrary, in machine learning systems the algorithm is extracted from the data in the sample set, as described above.[125] Mistakes will thus usually result from the input data and are more difficult to detect and fix. Hiding sensitive elements in the input date, such as ethnic background or geographical origin, could be considered in helping to prevent issues. However, even if those sensitive features are hidden, algorithms might nevertheless implicitly re-construct them from proxy variables.[126]

Moreover, as discussed above, the aim of machine learning is that the computer programs learn from experience and improve their performance over time.[127] The algorithm is therefore influenced not only by the original training dataset, but also by the use and continued data-input over time. Users therefore have a certain 'power' to change the algorithms. The swearing habit and other unacceptable behaviour of the AI-chatbot Tay, following interaction with its Twitter users, is a salient example.[128] One could also imagine that users in the legal context attempt to unduly influence or game the algorithms to obtain favourable results. For instance, if it were transparent that certain words, or cluster of words, such as in the study of the ECtHR decisions, led to a positive case prediction, the targeted use of those words in a party's legal submissions might lead to an inappropriate influence of the outcome.

Overall, this section has shown that a number of system-inherent limitations exist in the use of AI programs for legal decision-making. These limitations need to be carefully considered before promoting the use of AI in this context. Moreover, other more fundamental and wide-reaching concerns exist and are discussed in the next sections.

---

[124] Friedman and Nissenbaum describe a flight routing system sponsored by a US airline which systematically presented this airline on the first page. *See* Friedman & Nissenbaum, *supra* n. 114, at 331.

[125] *See supra* s. 2.

[126] Simon DeDeo, *Wrong Side of the Tracks: Big Data and Protected Categories* (2015), https://arxiv.org/pdf/1412.4643v2.pdf (accessed 9 May 2019) (for instance, income might be inferred from proxy variables such as postal codes).

[127] *See supra* s. 2.

[128] Ian Johnston, *AI Robots Learning Racism, Sexism and Other Prejudices from Humans, Study Finds*, The Independent (13 Apr. 2017), www.independent.co.uk/life-style/gadgets-and-tech/news/ai-robots-artificial-intelligence-racism-sexism-prejudice-bias-language-learn-from-humans-a7683161.html (accessed 9 May 2019) (Microsoft chatbot called Tay was given its own Twitter account and allowed to interact with the public; after twenty-four hours the chatbot used sexist, racist and profane language which it had learned from interaction with other Twitter users).

5 BLACK BOX OF LEGAL DECISION-MAKING WITH AI: NEED FOR REASONED DECISIONS

Providing a reasoned decision that outlines the premises on which it is based constitutes one of the fundamental features of legal decision-making. Schematically, one can distinguish several objectives for providing reasons in legal decisions. First, reasons help the losing party to understand why it lost and make the decision more acceptable (legitimacy objective). Second, reasons also allow the parties to the dispute, and if the decision is published, third parties in similar situations, to adapt their behaviour in the future (incentive objective). Third, reasons further allow other decision-makers to follow the same rationale or explain their departure therefrom (consistency objective). While one might discuss whether there is market for unreasoned decisions (e.g. in certain instances, parties might be interested in 'quick-and-dirty' unreasoned decisions), legal decisions must provide reasons unless the parties have provided otherwise.

AI programs will have significant issues in providing reasoned legal decisions and meeting those rationales.[129] Indeed, not only in the legal sector, but more broadly, the inability to explain results obtained with AI programs has raised concerns.[130] For example, disturbing results were obtained from an AI program able to guess a person's sexual orientation from publicly posted profile pictures.[131] The accuracy rates are troubling (83% for women and 91% for men) but what is even more alarming are the researchers' difficulties in determining the bases on which the AI program obtained those results.[132] This highlights the general problem for AI research of the so-called explainability or interpretability of its results.[133]

This difficulty is due to the features of certain AI models. Expert models or decision-trees follow pre-established rules, as detailed above.[134] It is therefore possible to identify the causes that led to a given result on the basis of those

---

129 *See* Scherer, *supra* n. 4, at 511–12.

130 *See* e.g. Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34:1 Berkeley Tech. L.J. 143 (2019).

131 Michal Kosinski & Yilun Wang, *Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images*, 114 J. Personality & Soc. Psychol. 246 (2018).

132 Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, New York Times (21 Nov. 2017), www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html (accessed 9 May 2019).

133 Or Biran & Courtenay Cotton, *Explanation and Justification in Machine Learning: A Survey*, in *IJCAI-17 Workshop on Explainable AI (XAI) Proceedings* 8 (2017), https://pdfs.semanticscholar.org/02e2/e79a77d8aabc1af1900ac80ceebac20abde4.pdf (accessed 9 May 2019) (defining interpretability as the ability for humans to understand operations either through introspection or through a produced explanation).

134 *See supra* s. 2.

rules and thus make the model explainable.[135] To the contrary, as also explained above, other machine learning models, such as neural networks, often have no pre-defined rules but use pattern-recognition to extract the required algorithm.[136] These systems may use hidden units which correspond to hidden attributes not directly observed.[137] As a consequence, the process by which those AI models obtain results is 'black-boxed' and not easily explainable.[138]

AI research tries to deal with those issues and develop Explainable Artificial Intelligence, also called XAI.[139] One possible route is the use of counterfactual scenarios. The model selects alternative samples with different features, compares the different outcomes under each and is therefore able to identify how and why they differ.[140] For instance, the model will be able to detect that the outcome in a given case would have been different, had feature X been absent or feature Y been added. In other words, the model for the actual decision-making is accompanied by another model, the purpose of which is to provide an explanation.[141]

The difficulty with providing reasoned legal decisions obtained by AI is two-fold. First, it may be difficult to identify the actual factors that have led to a certain outcome prediction in case of black-boxed models. Second, even if certain factors are identifiable as causes for a given outcome prediction, these factors might not prove a useful explanation. For instance, in the above-mentioned study on ECtHR decisions, certain words, or cluster of words, were identified with a high predictive value.[142] However,

---

135 *See* e.g. Bruce G. Buchanan & Edward H. Shortlie, *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley 1984).

136 *See supra* s. 2.

137 Alpaydin, *supra* n. 18, at 100.

138 *Ibid.*, at 155.

139 *See* earlier on Bruce Chandrasekaran, Michael C. Tanner & John R. Josephson, *Explaining Control Strategies in Problem Solving*, 4(1) IEEE Expert 9 (1989). *See* more recently Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 Harv. J.L. & Tech. 842 (2018). *See also* DARPA, *Explainable Artificial Intelligence (XAI) Program*, www.darpa.mil/program/explainable-artificial-intelligence (accessed 9 May 2019), full solicitation at www.darpa.mil/attachments/DARPA-BAA-16-53.pdf (2016) (accessed 9 May 2019); George Nott, *'Explainable Artificial Intelligence': Cracking Open the Black Box of AI*, Computer World (10 Apr. 2017), www.computerworld.com.au/article/617359/ (accessed 9 May 2019).

140 Charlotte S. Vlek et al., *A Method for Explaining Bayesian Networks for Legal Evidence with Scenarios*, 24 Artificial Intelligence L. 285 (2016).

141 *See* e.g. Michael Harradon, Jeff Druce & Brian Ruttenberg, *Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations* (2018), https://arxiv.org/abs/1802.00541 (accessed 9 May 2019); Bradley Hayes & Julie A. Shah, *Improving Robot Controller Transparency Through Autonomous Policy Explanation*, in *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction* (HRI 2017); Pat Langley et al., *Explainable Agency for Intelligent Autonomous Systems*, in *Proceedings of the Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence* 4762 (AAAI Press 2017); Marco T. Ribeiro, Sameer Singh & Carlos Guestrin, *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135 (ACM 2016).

142 *See supra* s. 3.

the information that words such as 'injury', 'Ukraine', 'copy', or 'January' have contributed to the outcome prediction falls short of an explanation which is deemed sufficient for a legally reasoned decision.

It is important to distinguish here between causal attribution, which is the process of extracting a causal chain and displaying it to a person, and causal explanation, which includes the social process of transferring knowledge between the explainer and the explainee with the goal that the explainee has the information needed to understand the causes of the event.[143] The latter not only requires AI to identify causes, but also to provide contextual explanation. Miller has shown that useful AI explanation must therefore take into account the human addressee.[144] This means, among other things, that explanation selection is important: typically, only a small subset of all possible causes are useful as an explanation for any given individual.[145] For instance, drawing from the ECtHR study results, the fact that an event happened in 'January' might be a cause for the decision, but less useful an explanation than that it constituted 'ill treatment'.

An explanation is also generally presented relative to the explainer's beliefs about the explainee's beliefs.[146] Dworkin has emphasized the importance of the shared context of law. In his major work, *Law's Empire*, he developed a theory of law as an interpretive practice that occurred in a community of interpreters.[147] Borrowing from the hermeneutical tradition, Dworkin claims that an understanding of a social practice, like law, requires turning to the meaning it has for participants. The meaning of law can therefore only be retrieved from within a shared context.[148] These contextual elements are likely to pose problems for AI-based legal explanation or reasoning.

Moreover, social scientists have tested the value of probabilistic explanations.[149] Overall, the use of statistical or probabilistic relationships are not as satisfying as causal explanations. For instance, if a student received a 50/100 in an exam and asks about the reasons for such score, the teacher's explanation that a majority of the class received the same score is unlikely to satisfy the student's request. Adding why most students received this score might be felt as an improvement, but not as much as explaining what this particular student did to receive his or her result.[150]

This example illustrates the difficulties for explanations or reasons in AI decision-making, which are, as detailed above, typically based on statistical or

---

143 Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, 267 Artificial Intelligence 1, at 17–18, 20 (2019).

144 *Ibid.*

145 *See* e.g. Denis J. Hilton, *Social Attribution and Explanation*, in *Oxford Handbook of Causal Reasoning* 645 (Michael Waldmann ed., Oxford University Press 2017).

146 *See* e.g. Denis J. Hilton, *Conversational Processes and Causal Explanation*, 107(1) Psychol. Bull. 65 (1990).

147 Ronald Dworkin, *Law's Empire* (Fontana 1986).

148 *Ibid.*

149 John R. Josephson & Susan G. Josephson, *Abductive Inference: Computation, Philosophy, Technology* (Cambridge University Press 1996).

150 Example provided by Miller, *supra* n. 143, para. 4.5.2.

probabilistic models.[151] Providing an 'explanation', say, that the likelihood of a claim to be dismissed is 86%, will not satisfy the losing party. It does not meet any of the objectives for legal reasoning outlined at the outset of this section. First, the legitimacy objective is not met, because statistical information is unlikely to help the losing party to understand why it lost and make the decision more acceptable. Second, the incentive objective fails because statistical information also does not allow parties or third parties to adapt their behaviour in the future. Finally, the consistency objective is not satisfied because other decision-makers have no information as to why they should follow the same rationale or depart therefrom.

The need for reasoned decisions is therefore likely to be an important barrier for AI-based legal decision-making. The impact of the probabilistic nature of AI models, however, raises even more fundamental questions as to the overall paradigm of decision-making, as discussed in the next section.

6 PARADIGM-SHIFT IN LEGAL DECISION MAKING: PROBABILISTIC INFERENCE INSTEAD OF DEDUCTIVE REASONING AND LOGIC?

Evaluating whether AI would be able to contribute to legal decision-making invariably raises the question how humans make legal decisions. As early as 1963, Lawlor speculated that computers would one day become able to analyse and predict judicial decisions, but noted that reliable prediction would depend on a 'scientific' understanding of the ways the law and the facts impact the judges' decision.[152] Even today, such 'scientific' understanding of judicial decision-making is lacking and is a debated topic amongst legal philosophers and theorists.

Theories of judicial decision-making abound, but a fundamental distinction exists between those that postulate the use of logic by ways of deductive reasoning on the basis of abstract, pre-determined legal rules (regrouped in the category of legal formalism), and those that emphasize the importance of extra-legal factors and the political dimension of the law (regrouped in the category of legal realism). This section shows that the use of AI in legal decision-making does not fit easily in either category. AI models would elevate probabilistic inferences to be the basis for legal decision-making and, as this section shows, this would constitute a sharp paradigm shift.

---

[151] *See supra* s. 2.

[152] Reed C. Lawlor, *What Computers Can Do: Analysis and Prediction of Judicial Decisions*, 49 ABA J. 337 (1963).

6.1 LEGAL FORMALISM AND THE USE OF DEDUCTIVE REASONING AND LOGIC

Legal formalism, in its purest form, posits that law is, and should be, an entirely self-contained system, in which judges never face choices or questions of interpretation that would be resolvable through extra-legal considerations.[153] Rather, as Max Weber put it, 'every concrete decision [is] the "application" of an abstract proposition to a concrete fact situation' and 'it must be possible in every concrete case to derive the decision from abstract legal propositions by means of legal logic'.[154]

A judicial decision is thus the product of a seemingly mechanical or mathematical application of pre-established legal principles or rules to the proven facts using means of logic.[155] The underlying idea can be expressed in the simple formula 'R + F = C' or 'rule plus facts yields conclusion'.[156] More specifically, the legal syllogism will consist of a major premise in the form of the pre-established rule (e.g. 'if P then Q') and a minor premise seeking to establish that the required condition stipulated in the major premise (P) occurred in fact. If such condition is met, by means of a deductive reasoning, or subsumption, the judge concludes that the legal consequence (Q) is to be applied in the case at hand as a matter of logic.[157]

Today, it is rare to find 'pure' formalists, but the main idea of legal decision-making as based on deductive reasoning and logic remains influential. In his seminal work *The Concept of Law*, Hart introduced an important distinction between clear cases, for which the simple deductive reasoning applies, and hard cases, for which extra-legal moral and political consideration may come into play.[158] Drawing on the philosophy of Wittgenstein, Hart emphasizes the indeterminacy of natural language and the open texture of law, for instance, through the use of general standards, such as 'good faith'.[159]

Even in their more nuanced forms, legal formalist theories still point to deductive, logical, rule-based reasoning as the guarantee for the objectivity, impartiality, and neutrality of law. MacCormick wrote in 1994:

---

153 *See* e.g. Hans Kelsen, *Reine Rechtslehre* 478 (2d ed., Deuticke 1960).

154 Max Weber, *Wirtschaft und Gesellschaft (Economy and Society)* 657–58 (Tübingen 1922).

155 French jurist Jean Domat saw the law as a logical, 'geometrical' demonstration, as any other scientific demonstration. *See* e.g. Marie-France Renoux-Zagamé, *La figure du juge chez Domat*, 39 Droits 35 (2004); Marie-France Renoux-Zagamé, *Domat, Jean*, in *Dictionnaire Historique des Juristes Français* (Patrick Arabeyre, Jean-Louis Halpérin & Jacques Krynen eds, Presses universitaires de France 2007).

156 Neil MacCormick, *Legal Reasoning and Legal Theory* x (Oxford Clarendon 1977) (with revised foreword, 1994).

157 *Ibid.*, at 21–29.

158 H. L. A. Hart, *The Concept of Law* (Oxford Clarendon 1961).

159 *Ibid.*

> A system of positive law, especially the law of a modern state, comprises an attempt to concretize broad principles of conduct in the form of relatively stable, clear, detailed and objectively comprehensible rules, and to provide an interpersonally trustworthy and acceptable process for putting these rules into effect. [...] [T]he logic of rule-application is the central logic of the law within the modern paradigm of legal rationality under the 'rule of law'.[160]

AI processes, if applied in the legal context, would potentially run counter to this understanding of legal decision-making. As described above in section 2, some computer models (such as expert models) are indeed rule-based, using causal logic and deductive reasoning, since they apply pre-established rules in the algorithm to the observable data. Other AI models, however, have different features. In particular, machine learning models, such as neural networks, often have no pre-defined rules. Deductive, causal reasoning is thus replaced by an inverse approach, because the machine learning program extracts the algorithm from the observable data. Rather than using logic, the AI model calculates probabilities, i.e. the likelihood for any given outcome.[161]

Applying such machine learning processes in the legal decision-making context therefore would mean accepting a departure from the above-mentioned understanding of judicial reasoning according to formalist theories. A decision based on those AI models would *not* be based on pre-determined legal rules, would *not* be the result of deductive logic, and would *not* follow the above-described legal syllogism. While this situation would be a cause for concern for legal formalists, it might be seen as vindicating others who have long criticized formalist theories.

6.2 Legal realism and the importance of extra-legal factors

Legal formalism has attracted important criticism over time. In the first half of the twentieth century, legal realists attacked the fundamental postulates of formalists theories.[162] Even though realist theories vary significantly, they have some commonalities. Llewelyn and others attacked the idea that the law was a mechanical application of pre-determined rules by the judge by means of logic and deductive reasoning.[163] Accepting that legal certainty was a myth, realists,

---

160 MacCormick, *supra* n. 156, at ix–x.

161 *See supra* s. 2.

162 For an overview *see* e.g. Laura Kalman, *Legal Realism at Yale: 1927–1960* (University of North Carolina Press, 1986); Wilfrid E. Rumble, Jr., *American Legal Realism: Skepticism, Reform and the Judicial Process* (Cornell University Press 1968). *See also* more recently Pierre Brunet, *Analyse Réaliste du Jugement Juridique*, 147:4 Cahiers Philosophiques 9 (2016); Brian Leiter, *Naturalizing Jurisprudence. Essays on American Legal Realism and Naturalism in Legal Philosophy* (Oxford University Press 2007).

163 *See* e.g. Karl N. Llewellyn, *Some Realism About Realism: Responding to Dean Pound*, 44(8) Harvard L. Rev. 1222 (1931). *See also* the later study, Wilfrid E. Rumble, Jr., *Rule-Skepticism and the Role of the Judge: A Study of American Legal Realism*, 15 Emory L.J. 251 (1966).

such as Frank, developed what they called rule scepticism and drew attention to the fact that rules do not play a determinative part in legal decision-making.[164] Rather, judges decide cases based on extraneous non-legal factors or their 'hunches' and then *ex post* provide their decision with a seemingly logical rule-deferring coating.[165] Unmasking the hypocrisy and double-standard of judicial decision-making, realists argue that logic and rule-deference is only a facade and ignores the social interests involved. This thought was later developed by the movement of critical legal theory, emphasizing the political significance of the law as a means of empowerment and emancipation.[166] Rather than being a mechanical and supposedly neutral application of rules, law does not contain a 'right answer' but corresponds to competing normative visions.[167]

Even before the legal realist movement became well-known, Justice Oliver Wendell Holmes described decision-making in similar ways. In 1897, in his seminal work, *The Path of Law*, he criticized what he called the 'fallacy of logic':

> certainty generally is illusion, and repose is not the destiny of man. Behind the logical form lies a judgment as to the relative worth and importance of competing legislative grounds, often an inarticulate and unconscious judgment, it is true, and yet the very root and nerve of the whole proceeding. You can give any conclusion a logical form.[168]

He insisted that law was imminently a matter of prediction, emphasizing the importance of statistics for the future of the law. He described his work as a study on prediction and more precisely 'the prediction of the incidence of the public force through the instrumentality of the courts'.[169] He thus argued that 'a legal duty so called is nothing but a prediction that if a man [or woman] does or omits certain things, he [or she] will be made to suffer in this or that way by judgment of the court; and so of a legal right'.[170] In order to make correct predictions, he surmised on the use of statistics for future lawyers' generations, noting that '[f]or the rational study of the law the black-letter man [or woman] may be the man [or woman] of the present, but the man [or woman] of the future

---

164 *See* e.g. Jerome Frank, *Law and the Modern Mind* (Brentano's 1930); Jerome Frank, *What Courts Do in Fact*, 26 Ill. L. Rev. 645, 645–66, 761–84 (1932).

165 Joseph C. Hutcheson, Jr., *The Judgment Intuitive: The Function of the 'Hunch' in Judicial Decision*, 14 Cornell L. Rev. 274 (1929).

166 *See* e.g. feminist critiques of adjudication, such as by Carol Gillian (e.g. *In a Different Voice* (Harvard University Press 1982)) and Catharine A. MacKinnon (e.g. *Feminism Unmodified: Discourses on Life and Law* (Harvard University Press 1987); *Toward a Feminist Theory of the State* (Harvard University Press 1989)).

167 *See* e.g. Roberto Mangabeira Unger, *The Critical Legal Studies Movement* (Harvard University Press 1983). Compare Antonin Scalia, *The Rule of Law as a Law of Rules*, 56 U. Chi. L. Rev. 1175 (1989) (arguing to reduce the discretion given to courts).

168 Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 Harv. L. Rev. 457, 466 (1897).

169 *Ibid.*, at 457.

170 *Ibid.*, at 458.

is [one] of statistics and the master of economics', adding that '[t]he number of our predictions when generalized and reduced to a system is not unmanageably large'.[171]

Holmes's emphasis in 1897 on prediction and statistics in legal decision-making, in lieu of logic, shines today in new light when considering the implications of AI. As discussed above, predictions based on statistics or probabilities are precisely features used in AI machine learning models.[172] Moreover, the importance of extraneous non-legal factors, as argued by the legal realists, is confirmed by the predictive AI studies, cited above.[173] In the ECtHR study, the part of the judgments with the highest predictive value is not the legal section but the section relating to the factual background.[174] Also, the US Supreme Court study included in the computer model extra-legal elements such as the judges' political preferences.[175]

Are we therefore to conclude, as some have argued,[176] that AI would vindicate the legal realists' theories? And that the possible use of machine learning models in legal decision-making would be in line with what human judges have always done? Would therefore, in essence, the debate between formalists and realists eventually be won by the latter? These conclusions, however, ignore an important point: the central place of probabilities as a normative basis for AI machine learning. As discussed in the next section, this goes well beyond legal realist theories.

6.3 Use of probabilistic inferences: Towards legal determinism?

When discussing legal theories on judicial decision-making, an important distinction needs to be drawn between their descriptive aspect (i.e. how judges *do* effectively reason and make decisions) and their prescriptive or normative aspect (i.e. how they *should* reason and make decisions).[177]

Legal formalism contains both a descriptive and normative element. Formalists *describe* the process by which judges apply the law as a matter of logic, deduction, and legal syllogism.[178] They also argue that the self-contained nature of the law, the neutrality of legal thinking untouched by extraneous non-legal factors is, normatively, how it *should*

---

171 *Ibid.*, at 458, 469.
172 *See supra* s. 2.
173 *See supra* s. 3.
174 Aletras et al., *supra* n. 45, at 10. *See supra* s. 3.1.
175 For a full list of the features, *see* Katz, Bommarito & Blackman, *supra* n. 69, at 4–6.
176 Aletras et al., *supra* n. 45, at 16 (who argued that their study results 'back … basic legal realist intuitions'). *See supra* s. 3.1.
177 *See* e.g. H. L. A. Hart, *Essays in Jurisprudence and Philosophy* 103–05 (Oxford Clarendon 1983). *See also* e.g. Pierre Brunet, *Le Raisonnement Juridique: Une Pratique Spécifique?* 26(4) Int'l J. Semiotics L. 767 (2013).
178 *See supra* s. 6.1.

be. This is in order to keep the law clear of politics or morality[179] and provide for a 'modern paradigm of legal rationality under the "rule of law"'.[180]

Legal realism, to the contrary, is first and foremost concerned with descriptive aspects. Holmes, Frank and others trace the *reality* of judicial decision-making – hence the name of the movement. They highlight the influence of extraneous non-legal factors, criticizing the formalistic, automatic, mathematical rule-application approach as utopian and far from the real world. However, they do not go as far as arguing that judges *should* take into account extraneous non-legal factors. To use the Israeli parole study, mentioned above,[181] as an illustration: while it might be a matter of fact that judges are influenced by extraneous factors such as food breaks, no one seriously argues that this is a good thing and should be the normative basis for judicial activity.

Normative aspects are not entirely foreign, though, to other theories, such as the critical legal theory movement, for instance. Unger and others have stressed the political significance of law and the social interests involved. Bringing out the normative aspects, law is taken as a means for effective radical social transformation.[182]

When looking at AI models, the foregoing leads to a number of observations. AI models would not only decide based on probabilities as a matter of fact, but would also be their normative basis. As mentioned above, a decision based on machine learning AI models would *not* be based on pre-determined legal rules, would *not* be the result of deductive logic, and would *not* follow the above-described legal syllogism.[183] This would be true on a descriptive level (i.e. how these models do effectively decide) and, importantly, also on a normative level (i.e. how these models should decide). Replacing logical, deductive and rule-based reasoning by probabilistic inferences as the normative framework of judicial decision-making would therefore not only constitute a departure from legal formalism, but would also go well beyond legal realists' theories.

Indeed, realists accept that judges, after having made their decision based on a variety of factors, including non-legal, political, and moral considerations, do render their decision coated in a format that seeks to comply with logic, using a rule-based deductive reasoning.[184] What realists criticize is the hypocrisies of such a facade, but they accept that such facade or format exists. AI-based decision-

---

179 *See* e.g. Kelsen, *supra* n. 153, at 478 ('What is here chiefly important is to liberate law from the associate which has traditionally been made for it – its association with morals.').

180 MacCormick, *supra* n. 156, at ix–x.

181 *See supra* s. 4.4.

182 *See* e.g. Unger, *supra* n. 167.

183 *See supra* s. 6.1.

184 *See* e.g. Holmes *supra* n. 168, at 465–66 ('The training of lawyers is a training in logic. The processes of analogy, discrimination, and deduction are those in which they are most at home. The language of judicial decision is mainly the language of logic. And the logical method and form flatter that longing for certainty and for repose which is in every human mind.').

making would take away such format. AI decisions would not be rendered making reference to deductive or causal reasoning based on legal rules. The problems related to this lack of reasoning have already been highlighted in section 5 above.

More fundamentally, however, the absence of a logical framework in judicial decision-making has implications that go beyond the descriptive or normative aspects discussed. Hart has distinguished three levels in judicial reasoning: (1) the processes or habits of thought by which judges actually reach their decision (descriptive psychology); (2) recommendations concerning the processes to be followed (prescriptive judicial technology); and (3) the standards by which judicial decisions are to be appraised.[185] It is at the third level that the absence of logic, at a minimum, causes concern because it undermines the assessment or justification of the decision. Or as Hart puts it:

> the issue is not one regarding the manner in which judges do, or should, come to their decisions; rather, it concerns the standards they respect in justifying decisions, however reached. The presence or absence of logic in the appraisal of decisions may be a reality whether the decisions are reached by calculation or by an intuitive leap.[186]

In addition, to the extent legal theories emphasize the political significance of law, as well as the fact that decision-makers have discretion to 'fill in' general standards, such as 'good faith',[187] the question arises how these political or moral considerations would be managed in an AI model. Who or what would be in a position to influence those political or moral considerations? In a traditional computer model, one might point towards the programmer. However, as described in section 2, in advanced AI models, the algorithm is not coded by a programmer but extracted from the observable data. Therefore, the only basis for the decision, even on morally or politically sensitive issues, will be past data. As already pointed out above, AI models are thus likely to take a conservative approach, even in a machine learning context of ever-improving algorithms.[188]

Using statistics or probabilities as the normative framework for judicial decision-making seems also problematic for other reasons. So far, probabilities or statistics are not an accepted legal basis for decisions.[189] English and other common law lawyers will be familiar with the term 'balance of probabilities' which sets out a standard of proof.[190] Importantly, however, this applies only to the establishment of facts. For instance, in *Miller v. Minister of Pensions*, the UK Supreme Court (then House of Lords) elaborated the balance of probabilities concept, stating that if 'the evidence is such that the tribunal can

---

185 Hart, *supra* n. 177, at 105.
186 *Ibid.*, at 105. *See also* Richard A. Wasserstrom, *The Judicial Decision* (Stanford University Press 1961).
187 *See supra* s. 6.1.
188 *See supra* s. 4.3.
189 *See* e.g. the discussion in the US Supreme Court case of *McCleskey v. Kemp*, 481 U.S. 279, 287 et seq. (1987).
190 *See* e.g. Emily Sherwin, *A Comparative View of Standards of Proof*, 50 Am. J. Comp. L. 243 (2002).

say "We think it is more probable than not", the burden is discharged, but if the probabilities are equal, then it is not'.[191] Once the facts are established using this method, probabilities have no room in judicial decision-making. For instance, one cannot grant a claim merely on the basis that there is an 80% chance that the established facts constitute a violation of the contract.

The previous example illustrates well the concrete issues with probabilistic bases for decision-making. What threshold would be appropriate above which a claim is deemed granted? Would anything above 50% be sufficient? Or would one require a higher threshold of, say, 80%? Even with such a higher threshold, though, one consciously accepts that there is a 20% likelihood that the case is decided wrongly.

In this context, it is also worth remembering the issue of data diet vulnerability and resulting bias risks, discussed above.[192] For instance, one might consider a situation where State X has been repeatedly found in violation of a substantive investment protection mechanism found in investment treaties. Does this influence the likelihood that State X will lose a future investment claim brought by another investor?

In sum, using a probabilistic analysis as the normative basis for decision-making is not only an important paradigm shift from a theoretical point of view, it also raises important concrete questions. This new approach could be called legal determinism since it determines future outcome on probabilistic calculations based on past data. As shown in this article, it has a number of implications for judicial decision-making which need to be considered carefully.

7 CONCLUSION

The aim of this article is to explore the use of AI in arbitral or judicial decision-making. Having assessed the technical aspects of AI and its implications and limitations, as well as the more fundamental impact it may have on human decision-making processes and theories thereof, the main findings and conclusions of this study are as follows:

- Existing studies on decision outcome prediction, while obtaining spectacular accuracy rates of 70–80%, contain important limitations. An analysis of the methodology and assumptions employed puts into doubt the claim these models might pave the way for *ex ante* outcome predictions. Among other things, it is questionable whether the models may equally apply and provide successful results for cases where the court originally decides a dispute, rather than reviewing a lower court's decision.[193]

---

191 House of Lords, [1947] 2 All E.R. 372 (opinion delivered by Lord Denning).
192 *See supra* s. 4.4.
193 *See supra* s. 3.

- The technical features of AI imply certain requirements for its use in judicial decision-making, at least as of today. This includes, for instance, the need for sufficient non-confidential case data[194] and, possibly, the requirement of repetitive fact-patterns and binary outcomes.[195] Given that AI models are typically based on information extracted from previous input data – even in ever-improving machine learning algorithms – they are likely to follow 'conservative' approaches and might not be adapted to deal with important policy changes over time.[196] Also, a blind deferential attitude towards algorithmic objectivity and infallibility is misplaced. Any data-based computer models are only as good as the input data, and there is therefore a risk that they perpetuate existing biases.[197]
- The need for reasoned decisions is likely to be an important barrier for AI-based legal decision-making. At least at the current technological level, it may be difficult to identify the factors that have led to a certain outcome prediction in case of black-boxed models. Moreover, even if certain factors are identifiable as causes for a given outcome prediction, these factors might not prove a useful explanation for human addressees in a given context.[198]
- The use of AI does not fit easily in legal theories on judicial decision-making. AI models elevate probabilistic inferences to be the normative basis for legal decision-making. This not only constitutes a paradigm shift from a theoretical point of view, but also raises important questions as to whether and how the outcome of future decisions should be determined on probabilistic calculations based on past data.

These conclusions, however, should not detract from the most obvious point: AI will fundamentally affect the legal profession and legal activities, including judicial decision-making. It is therefore important to study further how best to use AI, even with the limitations, barriers, and issues highlighted in this article. In international arbitration, which is under constant criticism for being too expensive and time-consuming, the claim by some AI developers that computers 'can do the work that took lawyers 360,000 hours' must be taken seriously. Future research is necessary to explore ways human decision-makers and AI can best be combined to obtain the most efficient results. Coming back to the quotation from Antoine de Saint-Exupéry in the introduction, we may not be able to foresee what the future of AI models looks like, but we can enable that future by carefully considering the implications of judicial decision-making with AI.

---

194 *See supra* s. 4.1.
195 *See supra* s. 4.2.
196 *See supra* s. 4.3.
197 *See supra* s. 4.4.
198 *See supra* s. 5.