Kafkaesque AI? Legal Decision-Making in the Era of Machine Learning

CAROLIN KEMPER*

ABSTRACT

Artificial Intelligence ("AI") is already being employed to make critical legal decisions in many countries all over the world. The use of AI in decision-making is a widely debated issue due to allegations of bias, opacity, and lack of accountability. For many, algorithmic decision-making seems obscure, inscrutable, or virtually dystopic. Like in Kafka's The Trial, the decision-makers are anonymous and cannot be challenged in a discursive manner. This article addresses the question of how AI technology can be used for legal decision-making and decision-support without appearing Kafkaesque.

First, two types of machine learning algorithms are outlined: both Decision Trees and Artificial Neural Networks are commonly used in decision-making software. The real-world use of those technologies is shown on a few examples. Three types of use-cases are identified, depending on how directly humans are influenced by the decision. To establish criteria for evaluating the use of AI in decision-making, machine ethics, the theory of procedural justice, the rule of law, and the principles of due process are consulted. Subsequently, transparency, fairness, accountability, the right to be heard and the right to notice, as well as dignity and respect are discussed. Furthermore, possible safeguards and potential solutions to tackle existing problems are presented. In conclusion, AI rendering decisions on humans does not have to be Kafkaesque. Many solutions and approaches offer possibilities to not only ameliorate the downsides of current AI technologies, but to enrich and enhance the legal system.

INTRODUCTION

> "Someone must have been telling lies about Josef K., he knew he had done nothing wrong but, one morning, he was arrested."[1]

In Franz Kafka's novel *The Trial*, Josef K. is arrested, prosecuted, sentenced, and, ultimately punished, without knowing the criminal charge, or meeting the prosecutor.[2] The arrest and the entire trial appear arbitrary and obscure; the way before the court is labyrinthine. The judges don't discuss his case with him. He is never confronted with inculpatory evidence. Legal authorities are portrayed as the epitome of a convoluted, opaque, inscrutable bureaucracy. *The Trial* (and Kafka's other works) coined the term *Kafkaesque*,

* Legal Trainee at the District Court of Mannheim (Germany); First State Exam, University of Mannheim (2018). The views expressed here are the author's own, and do not necessarily reflect those of any of the institutions with which she is affiliated.

1. FRANZ KAFKA, THE TRIAL 2 (David Wyllie trans., 2012) (1925).

2. *Id.*

used when something is "extremely unpleasant, frightening, and confusing,"[3] or "nightmarishly complex, bizarre, or illogical."[4] In such a *Kafkaesque* dystopia, citizens are subjected to an impersonal authority that does not explain or justify its actions. They cannot challenge the authority's decisions or hold the authority itself accountable.

In the wake of recent developments in the field of Artificial Intelligence ("AI"), parallels have been drawn between the use of AI and dystopias, such as Kafka's *The Trial*,[5] and George Orwell's *1984*.[6] Specifically, Machine Learning ("ML") algorithms have triggered discussion and alarm because of their "black box" nature and inherent opacity.[7] ML algorithms are used for a variety applications: facial recognition software[8] can identify suspects or wanted persons via public (or private) video surveillance.[9] Intelligent video surveillance can scan the behavior of people in hotspots to identify patterns indicating criminal offenses.[10] Lists with persons likely to be either victims or perpetrators of (gun) crimes are generated algorithmically.[11] Software can

3. *Kafkaesque*, CAMBRIDGE DICTIONARY, https://dictionary.cambridge.org/de/worterbuch/englisch/kafkaesque [https://perma.cc/RZC6-GJD7] (last visited Feb. 17, 2020).

4. *Kafkaesque*, DICTIONARY BY MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/Kafkaesque [https://perma.cc/M6T4-S23A] (last visited Feb. 17, 2020).

5. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1282 (2008); *see also* Gray Panthers v. S Schweiker, 652 F. 2d 146 ("Unless a person is adequately informed of the reasons for denial of a legal interest, a hearing serves no purpose and resembles more a scene from Kafka than a constitutional process.").

6. Jon Sharman, *Metropolitan Police's facial recognition technology 98% inaccurate, figures show*, INDEPENDENT (May 13, 2018), https://www.independent.co.uk/news/uk/home-news/met-police-facial-recognition-success-south-wales-trial-home-office-false-positive-a8345036.html13.05.2018 [https://perma.cc/54S7-U4LW].

7. *E.g.*, Bahar Gholipour, *We Need to Open the AI Black Box Before It's Too Late*, FUTURISM (Jan. 18, 2018), https://futurism.com/ai-bias-black-box [https://perma.cc/24KJ-PXPP]; Tom Simonite, *AI Experts Want to End "Black Box" Algorithms in Government*, WIRED (Oct. 18, 2017), https://www.wired.com/story/ai-experts-want-to-end-black-box-algorithms-in-government/ [https://perma.cc/9UUZ-4N9R]; *see also* Gary Marcus, *Deep Learning: A Critical Appraisal*, 10-11 (Jan. 2, 2018), https://arxiv.org/abs/1801.00631 [https://perma.cc/98M4-YFXZ]. *But see* Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 20 (Feb. 5, 2019), https://dx.doi.org/10.2139/ssrn.3329669 [https://perma.cc/HVN5-YA8X] (considering algorithms to be more transparent, because their decisions are traceable). *Cf.* Dallas Card, *The "Black Box" Metaphor in Machine Learning*, TOWARDS DATA SCIENCE (July 5, 2017), https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0 [https://perma.cc/7FD4-6DES].

8. A particularly sinister example is China's use of facial recognition technology to track the Uighur minority. *See* Paul Mozur, *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*, THE NEW YORK TIMES (April 14, 2019), https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html [https://perma.cc/YM9J-5S3U]. The potential misuse of facial recognition software has induced Google and Microsoft to restrain suppling their tools. *See* Brad Smith, *Facial Recognition: It's Time For Action*, THE OFFICIAL MICROSOFT BLOG (Dec. 6, 2018), https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/ [https://perma.cc/43JV-A4XS]; Kent Walker, *AI for Social Good in Asia Pacific*, GOOGLE BLOG (Dec. 13, 2018), https://www.blog.google/around-the-globe/google-asia/ai-social-good-asia-pacific/ [https://perma.cc/3PZR-E3T2].

9. On private video surveillance, *see* Alfred Ng, *With Facial Recognition, Shoplifting May Get You Banned in Places You've Never Been*, CNET (Mar. 20, 2019), https://www.cnet.com/news/with-facial-recognition-shoplifting-may-get-you-banned-in-places-youve-never-been/ [https://perma.cc/K4 W6-4BNA].

10. Amira Ben Mabrouk & Ezzeddine Zagrouba, *Abnormal Behavior Recognition for Intelligent Video Surveillance Systems: A Review*, 91 EXPERT SYST APPL 480 (2018); ALGORITHMWATCH, AUTOMATING SOCIETY—TAKING STOCK OF AUTOMATED DECISION-MAKING IN THE EU 82-83 (2019).

11. *Strategic Subject List*, CHICAGO DATA PORTAL, https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np [https://perma.cc/Y4DH-R2DJ]; Jeff Asher & Rob Arthur, *Inside the Algorithm That Tries to Predict Gun Violence in Chicago*, THE NEW YORK TIMES (Jun. 13, 2017),

calculate a recidivism score for offenders to determine their likelihood to reoffend, and, thereby, whether they will be granted parole.[12] Whether applicants are granted social benefits, and how much, is determined by software.[13] Some courts of Alternative or Online Dispute Resolution are automating their case management.[14] China is setting up a "social credit system"[15] using AI tools utilized by legal authorities in other countries.

Are these entities that (may) impinge on legal decisions, society, and, ultimately, our lives malefic or benign?[16] Or are they just conglomerations of statistical functions and data imprinted into the variables determining the outcome?[17] How will this technology impact and affect humans? These questions are of particular importance if state authorities deploy AI technologies: citizens might feel that they are at the mercy of an entity they do not understand and whose decisions are not transparent and explained. Citizens may perceive this as an illegitimate power gap between them and the state.[18] They may experience the Kafkaesque.

I. AI AND MACHINE LEARNING – IN A NUTSHELL

We typically associate terms like "mind", "consciousness", or human-like intelligence with "Artificial Intelligence". The current AI-technologies

https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html?searchResultPosition=1 [https://perma.cc/B5SE-493D].

12. *Compas Classification*, EQUIVANT, https://www.equivant.com/compas-classification/. [https://perma.cc/GZ99-ST29]; Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, THE NEW YORK TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html?searchResultPosition=2 [https://per ma.cc/9AXU-PEYW].

13. *E.g.*, in the United Kingdom, ALGORITHMWATCH, *supra* note 10, at 140-41.

14. Michael Legg, *The Future of Dispute Resolution: Online ADR and Online Courts*, 27 A.D.R.J. 227 (2016).

15. *See* Paul Mozur, *Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras*, THE NEW YORK TIMES (July 8, 2018), https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html [https://perma.cc/YMM2-6GYZ]; John Harris, *The Tyranny of Algorithms Is Part of Our Lives: Soon They Could Rate Everything We Do*, THE GUARDIAN (MAR. 5, 2018), https://www.theguardian.com/commentisfree/2018/mar/05/algorithms-rate-credit-scores-finances-data [https://perma.cc/ADJ3-BSV2]. *Cf.* Bing Song, *The West May Be Wrong About China's Social Credit System*, THE WASHINGTON POST (Nov. 29, 2018), https://www.washingtonpost.com/news/theworldpost/wp/2018/11/29/social-credit/?utm_term=.ec56ae3bf16e [https://perma.cc/Q3JL-QS2C] (suggesting that "social trust system" is the more appropriate term).

16. This insertion is only for vivid depiction of common concerns and worries. At this point, it does not make sense to anthropomorphize algorithms and software–there is no brain, and nothing is thinking. The impressive competence to solve certain tasks should not be misunderstood as competence. *Cf.* DANIEL C. DENNETT, FROM BACTERIA TO BACH AND BACK 56-75 (2017). Hence, algorithms do not have ulterior motives or hidden agendas (except the ones they were programmed or trained to achieve, especially by *reinforcement learning*).

17. Katarzyna Szymielewicz et al., *Black-Boxed Politics: Opacity Is a Choice in AI Systems*, MEDIUM (Jan. 17, 2020), https://medium.com/@szymielewicz/black-boxed-politics-cebc0d5a54ad [https://perma.cc/B64V-TYHL]; on the relationship between machine learning and statistics, *cf.* Joe Davison, *No, Machine Learning Is Not Just Glorified Statistics*, TOWARDS DATA SCIENCE (June 27, 2018), https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234 e3 [https://perma.cc/94RU-HZZ3].

18. LORENA JAUME-PALASÍ & MATTHIAS SPIELKAMP, ALGORITHMWATCH, ETHICS AND ALGORITHMIC PROCESSES FOR DECISION MAKING AND DECISION SUPPORT 14 (2017), http://algorithmwatch.org/wp-content/uploads/2017/06/AlgorithmWatch_Working-Paper_No_2_Ethics_ADM.pdf. [https://perma.cc/W9YJ-DN3H].

are related to mathematics, especially statistics, rather than reason and consciousness.[19] They are highly specialized to accurately perform one or several similar tasks and thus called "*Artificial Narrow Intelligence*".[20] An "*Artificial General Intelligence*", a human-level intelligence excelling in many different tasks, or even "*Artificial Superintelligence*" have yet to be achieved. Conversely, many algorithms and software tools once considered to be AI are no longer perceived as such, but rather as ordinary computation (dubbed the "*AI effect*" or Tesler's Theorem).[21]

A. THE BASICS OF MACHINE LEARNING

Lately, when we call something "AI", we mostly talk about ML algorithms or rather software containing one or several ML algorithms. There are different types of ML algorithms: decision trees, clustering algorithms, Artificial Neural Networks, and many more.[22] They mostly use conventional statistical methods (functions like linear or logistic regression, gradient descent, etc.) that approximate new data to existing examples in order to estimate the value of a target attribute.[23] During the learning process, ML algorithms typically use huge amounts of data to train and improve their results,[24] so that they can approximate new data more accurately.

19. *See* Oren Etzioni, *Deep Learning Isn't a Dangerous Magic Genie. It's Just Math*, WIRED (June 15, 2016), https://www.wired.com/2016/06/deep-learning-isnt-dangerous-magic-genie-just-math/ [https://perma.cc/4QS5-KNE5]; Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 95-100 (2014).

20. Recently, accomplishments in Transfer Learning were achieved. *E.g.*, Chrisantha Fernando et al., *PathNet: Evolution Channels Gradient Descent in Super Neural Networks* (Jan. 30, 2017), https://arxiv.org/abs/1701.08734 [https://perma.cc/YNC8-EGNQ]; *cf.* Albert Wu, *From Supervised Learning to Artificial General Intelligence*, MEDIUM (Dec. 12, 2018), https://medium.com/@albertwu_14963/from-supervised-learning-to-artificial-general-intelligence-196f0fbf601c [https://perma.cc/R3DS-LAAE].

21. *Cf.* Jennifer Kahn, *It's Alive!*, WIRED (Jan. 3, 2002), https://www.wired.com/2002/03/everywhere/ [https://perma.cc/FN7N-5PHY] (quoting Rodney Brooks: "Every time we figure out a piece of it, it stops being magical; we say, 'Oh, that's just a computation.'"). Tesler's Theorem states, "Intelligence is whatever machines haven't done yet." *See* Larry Tesler, *Tesler's Theorem, CV: Adages & Coinages*, http://www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html [https://perma.cc/ZA3U-QV36]. This is–according to Tesler himself–his original wording, not "Artificial Intelligence is whatever hasn't been done yet."

22. The following outline is based on ETHEM ALPAYDIN, MACHINE LEARNING: THE NEW AI (2016); IAN GOODFELLOW ET AL., DEEP LEARNING, 96-108 (2016); JOHN D. KELLEHER & BRENDAN TIERNEY, DATA SCIENCE 97-180 (2018); STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE 693-737 (2016). A comprehensive explanation of the process involved in developing ML algorithms can be found in David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669-702 (2017). Furthermore, Kleinberg et al., *supra* note 7, at 16-20 (clarifying that technically, there are two algorithms: one "model" or "screener"—the algorithm classifying or predicting based on the input—and the algorithm training the model by improving it through a learning process). As a quick-to-read and comprehensible introduction on ML, I recommend Gavin Edwards, *Machine Learning | An Introduction*, TOWARDS DATA SCIENCE (Nov. 18, 2018), https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0 [https://perma.cc/9652-NEXP] (last visited May 14, 2019); Vishal Maini, *Machine Learning for Humans*, MEDIUM (Aug. 19, 2017), https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12 [https://perma.cc/A52N-EK69].

23. ALPAYDIN, *supra* note 22, at 38-39; GOODFELLOW ET AL., *supra* note 22, at 96.

24. ALPAYDIN, *supra* note 22, at 104-08; GOODFELLOW ET AL., *supra* note 22, at 18-21; Edwards, *supra* note 22.

The learning process is different, depending on the type of the ML algorithm. There are three learning types: *supervised*, *unsupervised*, and *reinforcement learning*.[25] *Supervised learning* uses historic or manually processed data that is already classified (or *labeled*).[26] The algorithm will approximate its results to the available correct result contained in the *training data*.[27] *Unsupervised* learning trains an algorithm without labeled training data or any other explicit feedback.[28] Instead, it looks for regularities or hidden patterns by estimating the similarity of several data instances.[29] The algorithm re-executes with small adjustments until a certain quality target is reached. *Reinforcement learning* implements a continuous training process based on data which was obtained during operation.

The process of *supervised learning* consists of three steps.[30] In the *training phase*, the algorithm learns to perform predictions based on existing data to develop its topology. Usually, the algorithm randomly guesses the first topology and then adjusts by approximating to the given correct result.[31] During the *testing phase*, another dataset, the *test data*, will be provided to the algorithm.[32] The results on the test data will be compared to the correct result (or label) contained in the data sample.[33] The purpose of the testing phase is to find out how the ML algorithm performs on new data.[34] If it has perfectly learned the training data, it might be unable to cope with new data. This phenomenon is called *overfitting*, meaning the ML model only fits the training data, but performs poorly on new data.[35] Afterwards, the accuracy of several different algorithms will be compared in the *validation phase* to find the optimal algorithm.[36] The ML model can also be compared to human decision-making to determine whether it outperforms its alternatives. The training process and the quality of the training data determine the accuracy and robustness

25. Additionally, there is *semi-supervised learning*, in which the algorithm is partly trained with labeled (*supervised learning*) and partly with unlabeled data (*unsupervised learning*), OLIVIER CHAPELLE ET AL. EDS., SEMI-SUPERVISED LEARNING 2-3 (2006).

26. GOODFELLOW ET AL., *supra* note 22, at 103.

27. *Id.*

28. GOODFELLOW ET AL., *supra* note 22, at 103-04; RUSSELL & NORVIG, *supra* note 22, at 694-95.

29. ALPAYDIN, *supra* note 22, at 111-23; KELLEHER & TIERNEY, *supra* note 22, at 100-04; Maini, *supra* note 22.

30. For a more detailed account of the process with a focus on data science, *see* KELLEHER & TIERNEY, *supra* note 22, at 127-36, 145-48; Edwards, *supra* note 22.

31. KELLEHER & TIERNEY, *supra* note 22, at 129.

32. Lehr & Ohm, *supra* note 22, at 698-700.

33. KELLEHER & TIERNEY, *supra* note 22, at 145-47.

34. Lehr & Ohm, *supra* note 22, at 685-88, 698-700.

35. GOODFELLOW ET AL., *supra* note 22, at 108-18; Lehr & Ohm, *supra* note 22, at 684, 693-94; RUSSELL & NORVIG, *supra* note 22, 736-37.

36. KELLEHER & TIERNEY, *supra* note 22, at 147; Vishal Maini, *Machine Learning for Humans, Supervised Learning III*, MEDIUM (Aug. 19, 2017), https://medium.com/machine-learning-for-humans/supervised-learning-3-b1551b9c4930 [https://perma.cc/F3NS-YUDE].

of the ML algorithm.[37] Problems arise when the training data is biased, or the testing and cross-validation are not conducted thoroughly.[38]

B. DECISION TREES AND RANDOM FORESTS

One of the most common machine learning methods is the *Decision Tree*. It is used, for example, by the COMPAS Classification software that predicts recidivism,[39] or by credit bureaus that predict credit default. Decision Trees classify data and can make predictions by categorizing data with respect to expected results.[40] Typically, data is classified regarding a certain task, such as predicting a person's recidivism, by using pivotal features such as the criminal history, age, and sex of a defendant. The *root node* splits the data into two branches. (Does the defendant have a substantive criminal history: Yes or No.)[41] If the defendant has committed several offenses prior to his conviction the branch Yes might be selected, leading to the categorization as "most likely recidivistic" in a so-called *leaf node*, a node containing a final result. If the defendant had no criminal history prior to his conviction, the branch No will be activated. The next node (an *internal node*, which means branches lead to and from it) might split the data by regarding the defendant's compliance in prison.[42] If he was recalcitrant, the result "most likely recidivistic" would be obtained. If he was compliant, the next node, e.g. age or sex, would split the data, or another *leaf node* would come to the conclusion: "recidivism not likely."

37. Surden, *supra* note 19, at 106. Effects such as the Simpson's paradox—different data sets suggesting paradoxical conclusions; an eidetic explanation of Simpson's Paradox is contained in Lesser's poem *Confounded*—will cause distortions and inaccuracy. Lawrence M. Lesser, *Confounded*, 32 MATH. INTELL. 53 (2010).

38. GOODFELLOW ET AL., *supra* note 22, at 108-20; KELLEHER & TIERNEY, *supra* note 22, at 143.

39. *Classification Module*, EQUIVANT, https://www.equivant.com/compas-classification/ [https://perma.cc/TJ7B-ETDZ] (last visited June 2, 2020).

40. Carl Kingsford & Steven L. Salzberg, *What are Decision Trees?*, 26 NAT. BIOTECHNOL. 1011 (2008).

41. Disclaimer: This example is used for explanatory purposes only. Most decision trees with comparable tasks will be more complex and sophisticated (COMPAS uses a questionnaire with 137 questions; the responses are then used as features determining the classification). The example purports that recidivism will be presumed, and parole will be given or denied rigorously. This might not be the case in general, or with respect to different types of offenses.

42. Concerning compliance and recidivism, *see* Joshua C. Cochran & Daniel P. Mears, *The Path of Least Desistance: Inmate Compliance and Recidivism*, 34 JUSTICE QUARTERLY 431 (2017).

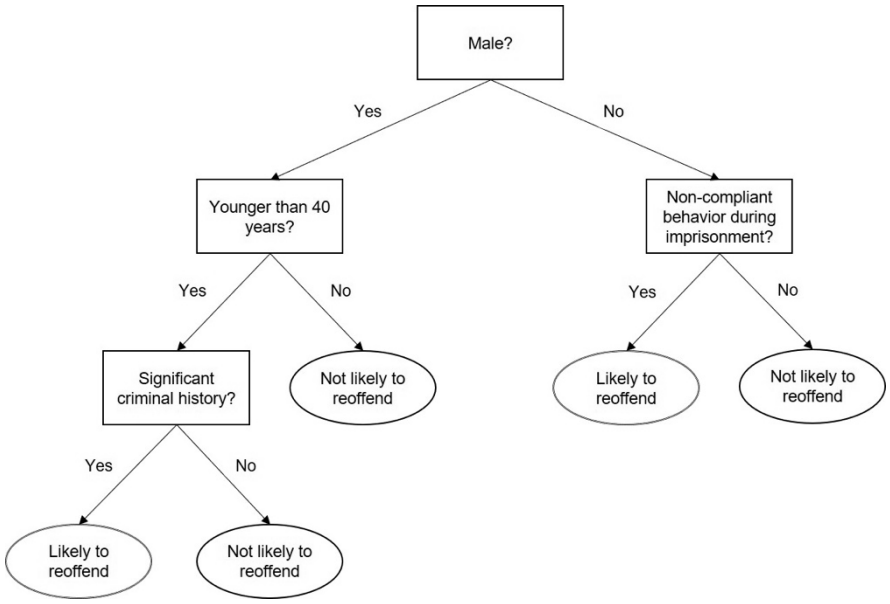

Figure 1: An exemplary Decision Tree predicting recidivism.

Decision Trees are simple and can be easily interpreted. To understand the reasoning process, one can follow the path through the branches. Why has Mr. X been classified to likely be recidivistic? He might not have a criminal history, but he has not been compliant during his imprisonment. However, in some cases the predictions of one Decision Tree may be not be very accurate. For example, people with prior convictions might not be recidivistic, or compliant inmates might still reoffend. To improve the accuracy of predictions by a Decision Tree, one can use several differently shaped Decisions Trees. The trees can be built up in a randomized way, meaning the first feature splitting the data at the Root will vary as well as the selection of the Internal Roots.[43] Based on a randomly selected sample of the training data, the tree's nodes are randomly determined. The different features split the data in a random order (e.g., 1) Sex, 2) Age, 3) Criminal History, …) and different ranges for variables will be set randomly (e.g., Age from below 20, 21 to 35, 35 to 60, above 60, or below 30, 31 to 33, 34 to 48, above 48). A multitude of trees will be designed this way by using different random data samples and different randomly constructed trees, until the random trees form a *Random Forest.* Now, if you enter new data (testing data or a new defendant), all your trees will classify whether recidivism is likely or not likely. They will cast "votes" congruent to their result, which will determine the Random Forest's result.

The predictions of Random Forests can achieve high levels of accuracy. The interpretability of their results is diminished, though. Instead of tracing

43. Leo Breiman, *Random Forests*, 45 MACHINE LEARNING 5 (2001); Maini, *supra* note 36.

the decision process of one tree, one needs to work through the results of a multitude of trees.[44]

C. ARTIFICIAL NEURAL NETWORKS AND DEEP LEARNING

Artificial Neural Networks ("ANNs") are much more complex than the first two methods: they involve several layers of "neurons" (which do not have anything to do with the human brain;[45] they are basically knots in a web-like structure and usually called *nodes*).[46] The input (typically data consisting of relevant measures and attributes) is received by the *input layer*.[47] The input nodes will forward the data to the next layer(s) which actually process the data – the so-called *hidden layers*. The *output layer* will render the "result" of the ANN. By applying mathematical methods of approximation towards the correct result in the hidden layers, the ANN "learns." More precisely, two nodes are linked via a *weight* that is modified during the learning process. The result of the ANN is determined by the entirety of all weights between nodes. Because of the many layers and the many adjusted weights in between, ANNs obtain their results in a very abstract manner.

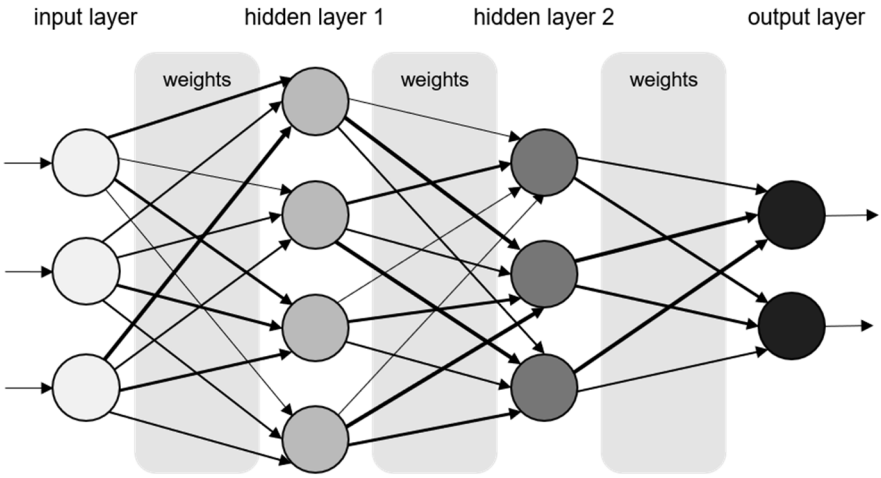

Figure 2: A schematic representation of an ANN with two hidden layers.

An ANN for object recognition (e.g., recognizing images with a cat) is trained via *supervised learning* with *labeled* training data, meaning the correct

44. Methods of interpretation of Random Forests are discussed by Prince Grover, *Intuitive Interpretation of Random Forest*, MEDIUM (Nov. 24, 2017), https://medium.com/usf-msds/intuitive-interpretation-of-random-forest-2238687cae45 [https://perma.cc/EBF9-X3J5]. One can look at the importance of different features (e.g. the importance of a criminal history) or use a Tree Interpreter to trace the prediction path of a particular case.

45. Neural Networks are named after the first rudimentary model, the "McCulloch-Pitts-Neuron," to which Dennett referred as "one of the great oversimplifications of all time." DENNETT, *supra* note 16, at 110.

46. GOODFELLOW ET AL., *supra* note 22, at 164-65; RUSSELL & NORVIG, *supra* note 22, 727-28.

47. KELLEHER & TIERNEY, *supra* note 22, at 123-24.

results (cat or no cat) are derived from historic data. In the beginning, the algorithm will frequently render a wrong result. For example, ANN misclassifies an image containing a cat as one without a cat. ANN's weights thereupon need to be adjusted, so that the correct result will be obtained in future runs.

Deep Learning refers to ANNs with more than two *hidden layers*.[48] These ANNs are typically trained with huge datasets (which is why *Big Data* is relevant for AI and machine learning). Deep Learning is employed when a use-case requires the consideration of many factors. It is used, for example, in image recognition to classify or segment depicted objects, such as GoogLeNet, with more than 20 layers. It can achieve impressive levels of accuracy. So-called *Convolutional Neural Networks* (CNN) can recognize handwritten characters with an accuracy of 99.7% (even humans are not be capable of attaining a 100%).[49]

As powerful as Deep Learning might be, it is also opaque: its process of obtaining the results is not intelligible due to the high levels of abstraction. What transpires in the hidden layers is unknown due to the high dimensionality of ML models. Hence, it is almost impossible to review or control the results.[50] Nevertheless, many AI researchers are dedicated to solving, or at least ameliorating, the transparency issue. Some approaches produce simplified approximations of the model as a whole (*global approximations*) or of an individual example (*local approximations*).[51] For example, *modular* ANNs can break open the process by disclosing intermediate steps leading to a result[52] and *probabilistic programming* (e.g., Google TensorFlow Probability) can provide information on the uncertainty of an ANN's prediction. Methods of explaining ML algorithms, especially ANNs, will be discussed in Part V.

II. EXAMPLES FOR AUTOMATED LEGAL DECISION-MAKING

In many countries, ML algorithms are used to assist with decision-making, or to automatically render a decision. The software used for automated decision-making ("ADM")[53] typically comprises many different components,

48. *Cf.* Jürgen Schmidhuber, *Deep Learning in Neural Networks: An Overview*, 61 NEURAL NETW 85 (2015); GOODFELLOW ET AL., *supra* note 22, at 164-67.

49. Yann LeCun et al., *The MNIST Database of Handwritten Digits*, http://yann.lecun.com/exdb/mnist/ [https://perma.cc/Q7QU-BWBU].

50. *See* ALPAYDIN, *supra* note 22, at 155-56.

51. The literature on Explaining AI/ML algorithms is quite extensive. A discussion of approximate models can be found in Brent Mittelstadt et al., *Explaining Explanations in AI*, PROCEEDINGS OF FAT* '19, 279, 281-83 (2019); Zachary C. Lipton, *The Mythos of Model Interpretability* (June 10, 2016), https://arxiv.org/abs/1606.03490 [https://perma.cc/JB39-3JZQ] (providing an overview on the topic).

52. *E.g.*, David Mascharka et al., *Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning* (Mar. 14, 2018), https://arxiv.org/abs/1803.05268 [https://perma.cc/72GH-D9VF].

53. Hereinafter, ADM can mean both automated and algorithmic decision-making. The two expressions are used interchangeably. Mostly, the term "ADM-software" will be used to not only denote the algorithms or the system rendering the decision, but the general software deployed for decision-making. A decision rendered by ADM-software will be referred to as an *algorithmic decision*.

functions, and algorithms. ML algorithms prevalently play a crucial role in determining an approximately optimal result.[54]

Software involved in legal decision-making can be categorized into *supportive technology* that assists the person making legal decisions, *replacement technology*, meaning tasks previously carried out by humans are now done by software, and *disruptive technology* that changes the way legal decision-makers work and ultimately reshapes our understanding of legal decision-making.[55] Another distinction can be made according to the systems autonomy: there is *decision support* (programs helping humans to make decisions), so called *human-in-the-loop* approaches (decisions are made with some human involvement) and completely autonomous *decision-making*.[56] Alternatively, one can categorize such software by assuming the perspective of the people affected by the decision. Some decision-support tools concern purely technical questions such as public procurement. Humans are either not at all or scarcely and only indirectly affected. Other decision-support tools provide information which directly influences the legal decision, for example, the likelihood of reoffending in parole proceedings. Lastly, a legal decision can be rendered by software itself, thus directly deciding upon humans. A system able to render legal advice or decision-making is called *Artificial Legal Intelligence*.[57] Hereinafter, examples of these three categories of legal decision-making will be portrayed.

A. PURELY TECHNICAL DECISION-SUPPORT TOOLS

Decision-support tools can help public administration and governments to address regulatory issues, such as government procurement,[58] or the allocation and planning of public services.[59] *Predictive Analytics* assists administrations with anticipating future needs.[60] Most of these decisions pertain to organizational efficiency but have no legal relevance (other than questions of

54. The specific algorithms used are not always known, since most of the software is proprietary and protected due to intellectual property rights or state secrets.

55. This taxonomy was developed by Tania Sourdin, *Judge v. Robot*, 41 UNSWLJ 1114, 1117 (2018).

56. Monika Zalnieriute et al., *The Rule of Law and Automation of Government Decision-Making*, 82 MODERN LAW REVIEW 1, 7 (pre-edited version 2019); Citron, *supra* note 5, at 1263-67 (distinguishing between fully automated systems and mixed systems where humans participate in the decision-making, review the decision, or provide further information or take over the case or request).

57. *Cf.* Sourdin, *supra* note 55, at 1122.

58. As done by the United Kingdom, ALGORITHMWATCH, *supra* note 10, at 142.

59. Ellen P. Goodman, *Defining Equity in Algorithmic Change*, THE REGULATORY REVIEW (Feb. 11, 2019), https://www.theregreview.org/2019/02/11/goodman-defining-equity-algorithmic-change/?etcc_med=newsletter&etcc_cmp=nl_algoethik_12522&etcc_plc=aufmacher&etcc_grp= [https://perma.cc/P8Z7-TV6L]. Examples such as carbon tax, congestion pricing, pollution allowances, and dynamic zoning codes are mentioned, as well as the optimization of bus routes in the United States. *Id.*

60. For example, the management of the German energy infrastructure is automated to ensure a more resilient supply and control of electricity. *See* Bundesnetzagentur (Federal Network Agency), *Security of Supply*, https://www.bundesnetzagentur.de/EN/Areas/Energy/Companies/SecurityOfSupply /SecuritySupply_node.html. [https://perma.cc/G92K-WKEU]. However, this increases the risk of cyber-attacks and enables surveillance via smart meters in private homes. ALGORITHMWATCH, *supra* note 10, at 82. Finland is planning to use AI to anticipate service needs and offering personalized services as well as digital assistants *Id.* at 60. In Denmark, ML algorithms predict the need of staff for geriatric care with 80 percent accuracy. *Id.* at 52.

compliance to budget plans and other administrative rules). Yet, some decision-support tools can have legal implications. Many countries use predictive policing systems that anticipate likely places or areas of criminal activity based on data of past crimes (*place-based predictive policing*).[61] Police presence will be increased in such areas during times when criminal activity is very likely. At first glance, no humans appear to be directly affected. However, these tools amplify police raids and patrols primarily in areas inhabited by minorities.[62] Some socio-economic, racial, or ethnic groups are more likely to be arrested than others.[63] As a consequence, they are more likely to be convicted. This in turn raises crime detection and consequently the crime rate of these areas, leading to more raids in the future, and therefore reinforcing the trend of policing the same areas (resulting in a *feedback loop*).[64]

B. DECISION-SUPPORT AFFECTING INDIVIDUALS

Some ADM-software assists legal authorities, such as public administrations, the police, and courts with decision-making. It can support legal authorities by recognizing certain features (e.g. a dialect indicating a person's origin, or the face of a wanted person) that may suggest a certain decision or action to be taken, or it may recommend decisions and actions. The final decision will be made by the competent judge or police officer. Yet, legal authorities may be inclined to abide by these suggestions. Thus, decision-support software has real consequences for citizens.

1. Recognition of Human Features: Dialect and Facial Recognition

In the wake of increasing refugee movement, Germany used dialect recognition software to determine the country of origin of an asylum seeker.[65]

61. *E.g.*, PREDPOL, https://www.predpol.com/ [https://perma.cc/BQ3B-JJE3]; ALGORITHM WATCH, *supra* note 10, at 44; *see also* Keith Kirkpatrick, *It's Not the Algorithm, It's the Data*, 60 COMM ACM 21, 22-23 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 129-37 (2017).

62. The Time Editorial Board, *The Problem with LAPD's Predictive Policing*, LOS ANGELES TIMES (Mar. 16, 2019), https://www.latimes.com/opinion/editorials/la-ed-lapd-predictive-policing-20190316-story.html [https://perma.cc/5Q24-M893].

63. *See, e.g.*, Radley Balko, *There's Overwhelming Evidence that the Criminal-Justice System Is Racist. Here's the pProof*, THE WASHINGTON POST (Sept. 18, 2018), https://www.washingtonpost.com/news/opinions/wp/2018/09/18/theres-overwhelming-evidence-that-the-criminal-justice-system-is-racist-heres-the-proof/ [https://perma.cc/UB9Z-BT6L]; Timothy Williams, *Black People Are Charged at a Higher Rate Than Whites. What if Prosecutors Didn't Know Their Race?*, THE NEW YORK TIMES (Jun. 12, 2019), https://www.nytimes.com/2019/06/12/us/prosecutor-race-blind-charging.html [https://perma.cc/K2DS-Y678].

64. Michael Veale et al., *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, PROCEEDINGS OF THE 2018 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1, 8 (2018); Selbst, *supra* note 61, at 135.

65. BAMF (FEDERAL OFFICE FOR MIGRATION AND REFUGEES), DIGITISATION AGENDA 2020 12 (2018), www.wir-sind-bund.de/SharedDocs/Anlagen/EN_nvam/Publikationen/Broschueren/broschuere-digitalisierungsagenda-2020.pdf;jsessionid=E9D95BA8CF50A1083E3E691E2EB75087.2_cid294?__blob=publicationFile [https://perma.cc/P3K9-4JGM] ("It provides an initial assessment, but it is no substitute for evaluation by staff of the Federal Office."); Amar Toor, *Germany to Use Voice Analysis Software to Help Determine Where Refugees Come From*, THE VERGE (Mar. 17, 2017), https://www.theverge.com/2017/3/17/14956532/germany-refugee-voice-analysis-dialect-speech-software [https://perma.cc/95KL-5DNG]; *Automatic Speech Analysis Software Used to Verify Refugees' Dialects*, DEUTSCHE WELLE (Mar. 17, 2017), https://www.dw.com/en/automatic-speech-analysis-software-used-to-verify-refugees-dialects/a-37980819?maca=en-tco-dw [https://perma.cc/346P-4T6K]; Justin Lee, *Germany to Use Voice Recognition Software to Analyze Refugees' Dialects*, BIOMETRIC UPDATE (Mar. 21, 2017),

This software was trained with recorded speech samples. The result was used as an indicator of the origin of the person seeking asylum. Since most asylum seekers arriving in Germany had no identification papers, the software facilitated determining a person's country of origin. This approach was met with doubts: the accuracy and reliability of the software was questioned. Unlike (human) linguistic experts assisting in the analysis of dialects, software may not be capable of taking into account the dynamic and changing nature of languages. Furthermore, the dataset used to train the software has to reflect changes in language and social conventions of speaking to different types of people.[66] Small ethnic groups and communities might not be adequately represented in these datasets.[67]

Facial recognition software is used in many countries to match the image of a person of interest with images stored in databases.[68] The threat of mass surveillance and the encroachment on democratic freedoms and people's privacy has led legislators in California to restrict the use of facial recognition software.[69] Tech companies like Google and Microsoft warned of the implications of their own technology and announced to limit access to facial recognition software.[70]

Another issue is the racial bias of several face recognition algorithms that come from training datasets where white faces are overrepresented.[71] Consequently, faces of individuals belonging to minority groups are not

https://www.biometricupdate.com/201703/germany-to-use-voice-recognition-software-to-analyze-refugees-dialects [https://perma.cc/N9PF-FVXD]; Ben Knight, *Germany 'Failed to use language recognition tech on refugees'*, DEUTSCHE WELLE (May 26, 2017), https://www.dw.com/en/germany-failed-to-use-language-recognition-tech-on-refugees/a-39001280 [https://perma.cc/P4JZ-HEC6].

66. People generally speak differently to their grandmother than to a government official, or to a friend. This argument was raised by Dirk Hovy, quoted in Philipp Hummel, *Software soll Dialekt von Asylbewerbern untersuchen*, WELT (Mar. 17, 2017), https://www.welt.de/wissenschaft/article162926845/Software-soll-Dialekt-von-Asylbewerbern-untersuchen.html [https://perma.cc/2LMV-9FJ6] (Ger.).

67. *Cf.* Judith Rosenhouse, *A Forensic Linguistics Problem*, 2 INT'L. J. LEGAL DISCOURSE 113 (2017) (identifying several under-documented Arabic dialects).

68. ALGORITHMWATCH, *supra* note 10, at 90, 140; Mozur, *supra* note 15. The Metropolitan Police in London recently deployed Live Facial Recognition technology on a trial basis. *See Live Facial Recognition*, METROPOLITAN POLICE, https://www.met.police.uk/live-facial-recognition-trial/ [https: //perma.cc/HPT6-NXZ7] (last visited Feb. 22, 2020). *Cf.* Bridges v. Chief Constable of South Wales, [2019] EWHC (Admin) 2341 (Eng.) (case against the South Wales Police in the United Kingdom concerning Automated Facial Recognition software).

69. San Francisco barred the police from using facial recognition software. *See* S.F. ADMIN, CODE § 19B.1-.8; Kate Conger et al., *San Francisco Bans Facial Recognition Technology*, THE NEW YORK TIMES (May 14, 2019), https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html [https://perma.cc/CKM6-CEE5]; *cf.* also the newly introduced Ordinance on Acquisition, Retention, and Use of Surveillance Technology, S.F. ADMIN, CODE § 19B.2, 19B.8, 19B.9 (2019). Additionally, California banned the use of facial recognition by police body cameras, CAL. PENAL CODE § 832.19 (Deering, 2019).

70. Smith, *supra* note 8; Walker, *supra* note 8.

71. Larry Hardesty, *Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems*, MIT NEWS (Feb. 11, 2018), http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 [https://perma.cc/366W-QYKF]; for a graphic presentation, *see* GENDER SHADES, http://gendershades.org/overview.html [https://perma.cc/8DPU-Q9NX] (last visited June 4, 2020). IBM is trying to reduce this bias by building a more diverse dataset. *Cf.* Devin Coldewey, *IBM Builds a More Diverse Million-Face Data Set to Help Reduce Bias in AI*, TECHCRUNCH (Jan. 29, 2019), https://techcrunch.com/2019/01/29/ibm-builds-a-more-diverse-million-face-dataset-to-help-reduce-bias-in-ai/ [https://perma.cc/G9LK-SN7E].

correctly recognized.[72] For example, faces of black individuals are mismatched or are not recognized at all.[73] Amazon's facial recognition software Rekognition obtained similarly bad results: it falsely recognized twenty eight members of Congress as having been arrested for crimes.[74] People of color were disproportionally represented (almost 40% of the false matches were people of color, although they constitute only 20% of Congress).

In general, many facial recognition systems simply do not perform well. The Metropolitan Police in London tested facial recognition software that scanned the faces of passers-by, checking whether they appear on databases containing individuals of interest.[75] Of 104 alerts, 102 were false positives[76] and only two were positive matches.[77]

2. Law and Order: Border Control, Predictive Policing, and Surveillance

Some ADM-software can discover and analyze terrorist networks by detecting terrorist-related online content and financial activities,[78] or analyze criminal behavior to identify trends and possible targets of sought-after suspects.[79]

A wide-ranging and versatile example of such software is *iBorderCtrl*, an "Intelligent Portable Border Control System",[80] a project funded by the European Union to speed up the process of border crossing. A pilot is about to be deployed at the borders of Hungary, Greece, and Latvia.[81] A *Biometrics*

72. Outrage ensued when algorithms recognized black people as gorillas. *Cf.* Tom Simonite, *When It Comes To Gorillas, Google Photos Remains Blind*, WIRED (Jan. 11, 2018), https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ [https://perma .cc/KU4M-EDXR]. The problem was "fixed." James Vincent, *Google 'Fixed' its Racist Algorithm by Removing Gorillas From Its Image-Labeling Tech*, THE VERGE (Jan. 12, 2018), https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai [https://perma. cc/E2KN-QX2G].

73. Tom Simonite, *The Best Algorithms Struggle to Recognize Black Faces Equally*, WIRED (Jul. 22, 2019), https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/ [https://perma.cc/A9KN-6YNZ]. A far more fatal example of racial bias are self-driving cars that detect pedestrians with darker skin tones less likely than light-skinned people. *See* Benjamin Wilson et al., *Predictive Inequity in Object Detection* (Feb. 21, 2019), https://arxiv.org/pdf/1902.11097 [https://per ma.cc/B2NU-79NJ].

74. Jacob Snow, *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*, ACLU (July 26, 2018), https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28 [https://perma.cc/C5BT-5C8X].

75. METROPOLITAN POLICE, *supra* note 68; *see also* ALGORITHMWATCH, *supra* note 10, at 140.

76. Sharman, *supra* note 6.

77. Sharman, *supra* note 6. The tool was dubbed "intrinsically Orwellian" by Big Brother Watch. *Id.*

78. The European Union uses the so-called DANTE ("Detecting and analysing terrorist-related online contents and financing activities"). ALGORITHMWATCH, *supra* note 10, at 35.

79. KeyCrime can even attribute a series of crimes to the same suspect. ALGORITHMWATCH, *supra* note 10, at 91.

80. *See Technical Framework*, IBORDERCTRL, https://www.iborderctrl.eu/Technical-Framework [https://perma.cc/6GZX-AYVU] (last visited Feb. 22, 2020); ALGORITHMWATCH, *supra* note 10, at 37-38. For criticism of the project, *cf.* Szymielewicz et al., *supra* note 17. Concerning ADM in Canada's immigration and refugee system, *cf.* PETRA MOLNAR & GILL LEX, U. TORONTO, BOTS AT THE GATE: A HUMAN RIGHTS ANALYSIS OF AUTOMATED DECISION-MAKING IN CANADA'S IMMIGRATION AND REFUGEE SYSTEM (2018).

81. IBORDERCTRL, https://www.iborderctrl.eu/Pilots [https://perma.cc/6SUH-HKLE] (last visited Feb. 22, 2020).

Module validates the biometric identity of the traveler. A *Face Matching Tool* creates a biometric signature that can be used to provide a matching score concerning future images of the traveler. A *Document Authenticity Analytics Tool* verifies travel documents and scans them for fraud characteristics. An *External Legacy and Social Interfaces System* crosschecks the traveler's information from social media or legacy systems like the already existing Schengen Information System. A *Risk Based Assessment Tool* aggregates and correlates collected data and estimates a risk score of a traveler, and thereby supports the decision-making of border guards. A *Hidden Human Detection Tool* detects people in various vehicles. An *Automatic Deception Detection System* performs, controls and assesses the pre-registration interview and calculates the probability of deceit in the interview by analyzing "non-verbal micro expressions".[82] The *Integrated Border Control Analytics Tool* helps to identify new patterns and knowledge and enables the iBorderCtrl system to quickly adapt to new situations. It also evaluates its performance and effectiveness, and analyses traffic data and can predict the expected traffic for certain dates.

Another – relatively novel – system is the "*Intelligent Video Surveillance System*" used in central squares and streets in Mannheim (Germany), to scan the behavior of people for patterns indicating "criminal offences such as hitting, running, kicking"[83] by means of automatic image processing. If such behavior is detected, police officers will be notified to review the respective video sequence and may take appropriate action. Instead of monitoring all video material, police officers will only see the video scenes of potentially criminal acts, while most of the footage will never be seen by anyone. Compared to other video surveillance systems, the encroachment on fundamental rights and the surveillance of citizens is reduced by this technology.

3. Individual-based Predictive Policing: A Grey Zone

The most invasive and truly "Orwellian" example of decision-support software is individual-based predictive policing. It is concerned with identifying possible future offenders and trying to take pre-emptive action. The most famous example is Chicago's *Strategic Subject List* (SSL).[84] The individuals on this list get assigned a score that reflects an "individual's probability of being involved in gun-violence either as a victim or an offender."[85] The features used to determine the score are: the number of times an individual was a victim of a shooting or aggravated battery or

82. This system is criticized in ALGORITHMWATCH, *supra* note 10, at 37, because of its average accuracy of 75% in relation to a single question. In IBORDERCTRL, *supra* note 80, the approach is defended by arguing that a 100% cannot be obtained and certain non-verbal behavior can indicate deception. The system itself generates a score that indicates the risk of deception, whereas the border guard can decide how to handle the situation and whether a second interview is needed (*human-in-the-loop principle*).

83. ALGORITHMWATCH, *supra* note 10, at 82; POLIZEIPRÄSIDIUM MANNHEIM, https://ppmannheim.polizei-bw.de/wp-content/uploads/sites/8/2018/09/VideoueberwachungMA_Info.pdf [https: //perma.cc/YV6S-UJDJ] (Eng. 4-5).

84. CHICAGO DATA PORTAL, *supra* note 11; *see also* Jessica Saunders et al., *Predictions Put Into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPER. CRIMINOL. 347 (2016).

85. CHICAGO DATA PORTAL, *supra* note 11.

assault, the individual's number of prior arrests, especially related to violent offenses, an individual's unlawful use of weapons or narcotics, an individual's trend in recent criminal activity and gang affiliation, and the age during an individual's most recent arrest.[86] The algorithm compiling the SSL aims to draw attention to a small group of people "who are at extreme levels of risk because their recent involvement in crime is so, so high".[87] The algorithm is continuously improved, yet the effectiveness of this strategy is so far questionable.[88]

One can justifiably argue that adding someone to the list is the result of an algorithmic decision about a human. Since the individual on the list will be subject to more policing, preventive measures, and social work, but will not be adjudicated, individual-based predictive policing is categorized as decision-support (severely) affecting humans. Nonetheless, the effects of these preventive actions can be intrusive, encroaching, degrading, and prejudicial.[89]

C. ALGORITHMIC DECISIONS ABOUT HUMANS

At the moment, ADM-software can barely render final decisions. Although adjudicating algorithms are envisaged, their development is still in the early stages. However, some decision-support tools induce legal decision-makers to rely on the algorithms' results. These ADM-tools consequently affect the decisions made concerning a human.

1. Assessing the Risk of Recidivism

Whether an offender may be released from prison on parole depends primarily on the estimated likelihood of him reoffending. To support judicial bodies in assessing the risk of recidivism, some countries, like the United States, use risk assessment software.[90] Such software typically uses Decision Trees or Random Forests to generate risk assessment scores.[91] Some judicial bodies use software assessing the risk or probability of reoffense to decide

86. Brianna Posadas, *How Strategic Is Chicago's "Strategic Subjects List"? Upturn Investigates*, MEDIUM (June 22, 2017), https://medium.com/equal-future/how-strategic-is-chicagos-strategic-subjects-list-upturn-investigates-9e5b4b235a7c [https://perma.cc/MR4X-2PDL]. By now, gang affiliation has been removed because of its weak impact on the score. *See* Josh Kaplan, *Predictive Policing and the Long Road to Transparency*, SOUTH SIDE WEEKLY (July 12, 2017), https://southsideweekly.com/predictive-policing-long-road-transparency/ [https://perma.cc/UU8D-LYBT].

87. Kaplan, *supra* note 86.

88. Saunders et al., *supra* note 84, at 366: ("The discrepancy between observed outcomes and predicted risk is operationally significant, but statistically reasonably small given the difficulty of predicting a low-probability event.").

89. For example, the individual on the SSL receives Custom Notifications, meaning the police visits them. Recipients are offered social services and informed about possible advanced penalties due to their criminal background. *Cf.* Kaplan, *supra* note 86.

90. Well-known to the brink of infamy is the tool COMPAS, used in the United States. *Cf.* EQUIVANT, *supra* note 39; Karen Hao, *AI Is Sending People to Jail—and Getting It Wrong*, MIT TECHNOLOGY REVIEW (Jan. 21, 2019), https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/?utm_campaign=site_visitor.unpaid.engagement&utm_medium=tr_social&utm_source=twitter [https://perma.cc/RT2F-77BM].

91. Janet Chan & Lyria Bennett Moses, *Is Big Data Challenging Criminology?*, 20 THEOR. CRIMINOL. 21, 31-32 (2016).

whether to parole a previous offender. Relevant details of a defendant, for example, his criminal history, his age, or his sex, are provided and processed to obtain a recidivism score. Software like the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS"), supplied by then-Northpointe (now Equivant), has been challenged by the public, who voiced allegations of racial bias.[92] In fact, people classified as "black" were deemed as likely to reoffend more frequently than those classified as "white."[93] The software provider defended COMPAS by arguing that black individuals were statistically more frequently recidivistic.[94] Hence, these statistics molded the software.

2. Adjudicating Algorithms and Artificial Legal Intelligence

Ordinary courts are still sparsely automated, but AI technology has already been introduced in some Alternative Dispute Resolution and Online Dispute Resolution institutions.[95] For example, judgements by default in debt collection proceedings are rendered solely by a program.[96] Estonia plans to have a "robot judge" that decides small claims disputes concerning contract law.[97] Such ADM-software typically comprises *Expert Systems* that consist of a knowledge base representing the knowledge and inference engines handling the reasoning. The e-Court system does not weigh arguments, or apply case law.[98] Humans, either as experts or as parties providing the relevant information through web forms, still play a major role.[99] ML models might eventually be harnessed to automate legal decision-making.[100] With *natural language processing*, the outcome of judicial proceedings can be predicted by analyzing legal materials, case law, or information on the judges.[101]

92. Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/4NU2-2J3S]; Angwin et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perm a.cc/R9V4-SPKT]; *cf.* Zalnieriute et al., *supra* note 56, at 11.

93. *See* Angwin et al., *Machine Bias*, *supra* note 92.

94. *Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity*, EQUIVANT, https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/ [https://perma.cc/C55M-C8CW].

95. Legg, *supra* note 14; Henriëtte Nakad-Westrate et al., *The Rise of the Robotic Judge in Modern Court Proceedings*, INT'L CONF. INFO. TECH. 59 (2015) (discussing the Dutch e-Court, rendering arbitrational verdicts).

96. *See, e.g.*, Nakad-Westrate et al., *supra* note 95.

97. Eric Niiler, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, WIRED (Mar. 25, 2019), https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ [https://perma.cc/34NY-6T74].

98. Nakad-Westrate et al., *supra* note 95.

99. Davide Carneiro et al., *Online Dispute Resolution: An Artificial Intelligence Perspective*, 41 ARTIFICIAL INTELL. REV. 211, 238 (2014).

100. Benjamin Alarie et al., *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L. J. 106, 115-16 (2018).

101. France recently banned analyzing decisions of individual judges. *Cf.* Malcolm Langford & Mikael Rask Madsen, *France Criminalises Research on Judges*, VERFASSUNGSBLOG (Jun. 22, 2019), https://verfassungsblog.de/france-criminalises-research-on-judges/ [https://perma.cc/85MT-DHSL]. *See also* KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS 107-26 (2017); N. B. Chaphalkar et al., *Prediction of Outcome of Construction Dispute Claims Using Multilayer Perceptron Neural Network Model*, 33 INT'L J. PROJ. MGMT. 1827 (2015); Nikolaos Aletras et al., *Predicting Judicial Decisions of the European Court of Human Rights: a Natural Language Processing Perspective*, 2 PEER J. COMPUTER SCI. 93 (2016). Other software

III. THEORETICAL FRAMEWORKS FOR ASSESSING ALGORITHMIC DECISION-MAKING OF LEGAL AUTHORITIES

Algorithmic decision-making can have direct and potentially far-reaching consequences for the concerned. These consequences will not differ from the consequences incurred if humans render decisions.[102] Even so, will citizens accept algorithmic decisions, or will they spurn nonhuman decisions? Which prerequisites need to be met to foster the acceptance of algorithmic decisions? These questions are pressing, demonstrated by the many forthright forewarnings of a dystopian and cataclysmic future voiced in the discourse about AI.[103] Indeed, decisions rendered by opaque, inscrutable computer programs might be perceived as Kafkaesque by concerned citizens.

Decisions of legal authorities need to be legitimate, inducing citizens to comply.[104] If social acceptance of algorithmic decision-making is to be enhanced, a criteria for evaluating the legitimacy of legal decisions must be established. In the following section, three theoretical frameworks aimed at promoting legitimacy and social acceptance of decision-making by legal authorities will be outlined. In the field of *machine ethics*, guidelines have been compiled on how a "good" algorithm should be developed and deployed. The main aspects are fairness, accountability, and transparency. The theory of *procedural justice* examines how citizens interact with legal authorities and establishes criteria that enhance the perceived legitimacy of authorities and their decisions. The constitutional principles of the *rule of law* and *due process* can provide valuable insights on fostering the legitimacy of decisions made by state authorities.

that is usually considered Artificial Legal Intelligence does not adjudicate but performs preliminary tasks such as document discovery. *See* Sourdin, *supra* note 55, at 1118.

102. A set of lawful consequences can be built into the ADM-software to prevent evident transgressions. One can assume that every possible output could at least hypothetically correspond to a conceivable human decision.

103. Stephen Hawking warns: "Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all." Abigail Higgins, *Stephen Hawking's Final Warning for Humanity: AI Is Coming for Us*, VOX (Oct. 16, 2018), https://www.vox.com/future-perfect/2018/10/16/17978596/stephen-hawking-ai-climate-change-robots-future-universe-earth [https://perma.cc/UL27-MGVN]. Elon Musk sees AI as the "biggest existential threat." Kelsey Piper, *Why Elon Musk Fears Artificial Intelligence*, VOX (Nov. 2, 2018), https://www.vox.com/future-perfect/2018/11/2/18053418/elon-musk-artificial-intelligence-google-deepmind-openai [https://perma.cc/5GKK-F3UH]. Additionally, AI researchers warn for safety concerns with Artificial Narrow Intelligence like ML. Dario Amodei et al., *Concrete Problems in AI Safety* 21 (June 21, 2016), https://arxiv.org/abs/1606.06565 [https://perma.cc/2M64-6DDD] ("While many current-day safety problems can and have been handled with ad hoc fixes or case-by-case rules, we believe that the increasing trend towards end-to-end, fully autonomous systems points towards the need for a unified approach to prevent these systems from causing unintended harm."); *see also Open Letter on Research Priorities for Robust and Beneficial Artificial Intelligence*, FUTURE OF LIFE INSTITUTE, https://futureoflife.org/ai-open-letter/?cn-reloaded=1 [https://perma.cc/YM47-QZPW].

104. Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME AND JUSTICE 283, at 307-10 (2003).

A. MACHINE ETHICS

Fairness, accountability, and transparency are imperatives for developing ethical ADM software.[105] Fairness can be assessed by searching for *disparate treatment* (disadvantaging the members of a groups, e.g., based on race or sex), *disparate impact* (meaning a practice has a disproportionate adverse effect on members of protected groups), or the impeding of *fair representation* of certain groups.[106] Transparency is concerned with disclosing and explaining the results in understandable terms.[107] Accountability refers to the problem that algorithms are not responsible for their own decisions, but the institutions deploying algorithms are. They should use ADM-software responsibly and in accordance with their mandate.[108]

The European Commission had a high-level expert group develop Ethics Guidelines for Trustworthy AI.[109] Four ethical principles are established in these guidelines: the principle of respect for human autonomy (meaning that humans interacting with AI must maintain full self-determination over themselves), the principle of prevention of harm, the principle of fairness, and the principle of explicability.[110] To achieve trustworthy AI, proper oversight is promoted, possibly through human-in-

105. An annual conference addresses these issues: ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*), https://fatconference.org/ [https://perma.cc/NX7M-4K46]; *e.g.*, referred to by Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-making Processes*, 31 PHILOS. TECH. 611 (2018). By now, many institutions have developed AI principles. *See, e.g.*, Press Release, IEEE, IEEE Ethically Aligned Design Document Elevates the Importance of Ethics in the Development of Artificial Intelligence and Autonomous Systems (Dec. 13, 2016), https://standards.ieee.org/news/2016/ethically_aligned_design.html [https://perma.cc/YG72-9ACC]; *Asilomar AI Principles*, The FUTURE OF LIFE INSTITUTE, https://futureoflife.org/ai-principles/ [https://perma.cc/Q3LE-GX9D] (distinguishing between *failure* and *judicial transparency*, but also listing, inter alia, responsibility, privacy, and human control). *See also* the somewhat similar principles of the *OECD Principles on AI*, ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (2019), http://www.oecd.org/going-digital/ai/principles/ [https://perma.cc/3C9N-BMBZ].

106. Kleinberg et al., *supra* note 7, at 6-7; Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* (Aug. 14, 2018), https://arxiv.org/abs/1808.00023 [https://perma.cc/89VW-UADF]; Alexandra Chouldechova & Aaron Roth, *The Frontiers of Fairness in Machine Learning* (Oct. 20, 2018), https://arxiv.org/abs/ 1810.08810 [https://perma.cc/32G2-DEDR]. Fairness can also be examined as *group* v. *individual fairness*, or equality of opportunity. *See* Lepri et al., *supra* note 105, at 614-18.

107. *See, e.g.*, on "interpretability", Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation* 2-3 (Nov. 3, 2017), https://arxiv.org/abs/1711.01134 [https://perma.cc/ 6ULL-G9D8]; Lepri et al., *supra* note 105, at 619-22.

108. *See, e.g.*, Nicholas Diakopoulos, *Accountability in Algorithmic Decision Making*, 59 COMM. ACM 56, 58 (2016) (explaining that software engineers should be mindful of being accountable and responsible for their work and avoid harm to others).

109. High-Level Expert Group on Artificial Intelligence (AI HLEG), European Commission, *Ethics Guidelines for Trustworthy AI* 2 (Apr. 8, 2019); *see also* Nathalie A. Smuha, *The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence*, COMPUTER L. REV. INT'L 97 (2019); Ramak Molavi Vasse'i, *The Ethical Guidelines for Trustworthy AI – A Procrastination of Effective Law Enforcement*, COMPUTER L. REV. INT'L 129 (2019).

110. *See* AI HLEG, *supra* note 109, at 11-13 (listing the principles of human autonomy, prevention of harm, fairness and explicability); *see also* Luciano Floridi et al., *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 MINDS MACH. 689, 695-700 (2018) (describing the Principle of Beneficence ("do good"), the Principle of Maleficence ("do no harm"), the Principle of Autonomy ("preserve human agency," meaning humans must keep full self-determination over themselves), the Principle of Justice ("be fair"), and the Principle of Explicability ("operate transparently"). These principles were established in the AI HLEG, *Draft Ethics Guidelines for Trustworthy AI* 8-10 (2018)).

the-loop, human-on-the-loop, and human-in-command approaches.[111] While a human-in-the-loop can intervene at any point, a human-on-the-loop merely monitors the ADM-software during the decision-making process, whereas a human-in-command oversees the deployment and use of ADM-software in general.[112] Transparency and explicability as well as technical robustness and safety are seen as crucial to build trust in AI systems.[113] According to the guidelines, privacy, data governance, diversity, non-discrimination, and fairness need to be observed.[114] Furthermore, the consideration of the environment and social and societal impact of AI systems is emphasized.[115] Lastly, accountability and responsibility of AI systems, for example, by auditing and accessible redress, are demanded.[116]

It is not just the European Union that grappled with the emergence of AI; the Council of Europe and its European Commission of the Efficiency of Justice adopted a framework for policy makers, legislators, and justice professionals concerning the use of AI in judicial systems.[117] The charter comprises five principles: the principle of respect for fundamental rights, the principle of non-discrimination, the principle of quality and security, the principle of transparency, impartiality and fairness, and the principle "under user control." Most principles concur with those found within the Ethics Guidelines for Trustworthy AI of the AI HLEG. For example, the aim of principle "under user control" is to ensure that users are informed actors and in control of their choices,[118] corresponding to the principle of human autonomy within the AI HLEG Ethics Guidelines. Specifically, decision-makers should be able to review the resultsof ADM-software and the relevant data used to produce these results.[119]

B. THE THEORY OF PROCEDURAL JUSTICE

When algorithms decide upon citizens, the affected individuals might question the legitimacy of the algorithm. They might reject or defy decisions they consider to be illegitimate and rebel against their enforcement. Algorithmic decisions need to ascertain legitimacy to promote compliance of the people involved. A rule or an authority has legitimacy "when others feel obligated to defer voluntarily."[120] According to the theory of Procedural Justice, compliance with decisions of legal authorities and the evaluation of

111. AI HLEG, *supra* note note 109, at 15-16.

112. *Id.* at 16.

113. *Id.* at 16-18.

114. *Id.* at 17-19.

115. *Id.* at 19.

116. *Id.* at 19-20, 26-31 (including an assessment list addressing the requirements of trustworthy AI).

117. *CEPEJ European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment*, EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE (Dec. 2018), https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment [https://perma.cc/Q7HD-US78].

118. *Id.* at 12.

119. Issues pertaining to human autonomy and user control with regard to human decision-makers will not be addressed here.

120. Tyler, *supra* note 104, at 307.

legitimacy depends on experienced fairness in the decision-making process.[121] The perception of fairness in a decision-making process, however, is not primarily based on the outcome, but instead based on the manner in which the process is handled.[122] Adverse decisions are accepted more willingly if the affected persons perceived the decision-making process as fair.[123] Legitimacy leads to normative compliance (internal values leading to voluntary deference to the law and to the decisions of legal authorities) instead of instrumental compliance (which is based on the favorability of a decision).[124] People do not evaluate actions and decisions of legal authorities by considering legality.[125]

Additionally, the evaluation of legal authorities involved—courts, judges, the police, or the law itself—depends on the experienced fairness.[126] Determinant aspects of procedural justice are the quality of decision-making itself and the quality of treatment during the process of decision-making.[127] While evaluating fairness, people do not only consider aspects that affect the decision-making itself, but also the treatment they experience.[128] Principles of procedural justice include voice, neutrality, respect, and trust:[129] People

121. *Cf.* extensive research done by Tyler and others, especially: Tom R. Tyler, *Enhancing Police Legitimacy*, 593 ANN. AM. ACAD. POLITICAL SOC. SCI. 84 (2004); Tyler, *supra* note 104; Tom R. Tyler, *Psychological Perspectives on Legitimacy and Legitimation*, 57 ANNU. REV. PSYCHOL. 375 (2006); TOM R. TYLER, WHY PEOPLE OBEY THE LAW (2006); Tom R. Tyler, *Procedural Justice and the Courts*, 44 COURT REV. 26 (2007); Steven L. Blader & Tom. R. Tyler, *A Four-Component Model of Procedural Justice: Defining the Meaning of a "Fair" Process*, 29 PERS. & SOC. PSYCHOL. BULL. 747 (2003); Tom R. Tyler & Cheryl J. Wakslak, *Profiling and Police Legitimacy: Procedural Justice, Attributions Of Motive, and Acceptance of Police Authority*, 42 CRIMINOLOGY 253 (2004); Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181, 276-277 (2004); Mike Hough et al., *Procedural Justice, Trust, and Institutional Legitimacy*, 4 POLICING 203 (2010); J. Jackson et al., *Why do People Comply with the Law?: Legitimacy and the Influence of Legal Institutions*, 52 BRITISH J. CRIMINOL. 1051 (2012); Lorraine Mazerolle et al., *Shaping Citizen Perceptions of Police Legitimacy: A Randomized Field Trial of Procedural Justice*, 51 CRIMINOLOGY 33 (2013); Tracey L. Meares et al., *Lawful or Fair? How Cops And Laypeople Perceive Good Policing*, 105 J. CRIM. L. & CRIMINOL. 297 (2015); Tom R. Tyler et al., *The Impact of Psychological Science on Policing in the United States: Procedural Justice, Legitimacy, and Effective Law Enforcement*, 16 PSYCHOL. SCI. PUBLIC INTEREST 75 (2015); Liesbeth Hulst et al., *On Why Procedural Justice Matters in Court Hearings: Experimental Evidence that Behavioral Disinhibition Weakens the Association between Procedural Justice and Evaluations of Judges*, 13 UTRECHT L. REV. 114 (2017).

122. Tyler, *supra* note 104, at 300-01; TYLER, WHY PEOPLE OBEY THE LAW, *supra* note 121, at 71-84; Tyler, *Procedural Justice and the Courts*, *supra* note 121, at 26-30; Solum, *supra* note 121, at 242-73 (contrasting accuracy of the decisions with the costs of adjudication and participation).

123. Tyler, *Procedural Justice and the Courts*, *supra* note 121, at 26-30; Tom R. Tyler, *Public Trust and Confidence in Legal Authorities: What Do Majority and Minority Group Members Wants from the Law and Legal Institutions*, 19 BEHAV. SCI. L. 215-16 (2001).

124. Citizens comply because of *value-based motivation*. TYLER, WHY PEOPLE OBEY THE LAW, *supra* note 121, at 115-24; Jason Sunshine & Tom R. Tyler, *The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing*, 37 LAW SOC. REV. 513, 517 (2003).

125. Meares et al., *supra* note 121, at 304.

126. Tyler, *Procedural Justice and the Courts*, *supra* note 121, at 26.

127. *See* Blader & Tyler, *supra* note 121, at 748-49 (explaining that the qualities of decision-making and treatment may be subsumed under the term procedural function). The authors further explain another dimension of decision-making, procedural source, which is either formal or informal influences on the decision-making procedure. *Id.* They consequently developed a four-component model by combining formal and informal sources with decision-making and treatment, thus obtaining four types of concern. *Id.*

128. Blader & Tyler, *supra* note 121, at 755.

129. Tyler, *supra* note 104, at 298-99; Tyler, *Procedural Justice and the Courts*, *supra* note 121, at 30-31; Tyler, *Enhancing Police Legitimacy*, *supra* note 121, at 94-95; Tyler et al., *The Impact of Psychological Science on Policing in the United States: Procedural Justice, Legitimacy, and Effective Law Enforcement*, *supra* note 121, at 85-86; Meares et al., *supra* note 121, at 307-11.

want an opportunity to tell their story; they want to voice their arguments, views, and needs. They wish for authorities to consider of their view and perspective before deciding. Solum considers participation of the parties concerned to be paramount in the perception of fairness in his theory of procedural justice.[130] Legal authorities are expected to be neutral and to apply the law consistently regardless of the case at hand and the individuals involved. If decision-makers are transparent and open about their decision-making, disclose the applied legal rules, and explain the reasons of a specific decision, they foster the perception of neutrality, objectivity and factuality. Respectful treatment of parties concerned matters at all stages of a decision-making process. It encompasses politeness, the respecting of people's rights, and treating them with dignity.[131] The people affected by the decision-making process should be able to participate in the process as autonomous and equal citizens.[132] Lastly, trust in legal authorities hinges on people trusting the motives of legal authorities.[133] If decision-makers appear to be sincere, well-intentioned, caring, and open about the reasons of their decisions, people infer that they listen, consider the views voiced, and act benevolently.[134]

A *process-based approach* realizes procedural justice principles can improve the success of policing: people will be more willing to comply with decisions and directives of the police.[135] Particular policing strategies can impair the perception of fairness and legitimacy of police action. Racial profiling will induce affected people to perceive the police as acting unfairly.[136] People generally do not feel responsible or accountable for characteristics that are beyond their control (such as race, or sex). Rather, they accept decisions based on their actions. If actions of authorities, such as the police, appear to be motivated by characteristics ascribed to them, they will refuse to assume responsibility. Instead they will infer that the police's actions are motivated by ascribed characteristics, such as race.[137] Consequently, they will not trust the authorities' motives for certain actions or decisions. But profiling does not only aggravate the direct subject of profiling actions. A study conducted by Tom R. Tyler and Cheryl J. Wakslak on procedural justice and profiling suggests that in general, people's support for the police depends on their

130. Solum, *supra* note 121, at 273-305. *See id.* at 267-68, for an explanation that trials are analyzed as the model of the ideal communication situation articulated by Jürgen Habermas in BETWEEN FACTS AND NORMS (1996). Each party capable of engaging in communication is allowed to participate and shall be given equal opportunity to communicate, especially questioning other participants' proposals, introducing new proposals, and expressing attitudes, beliefs, wishes, and needs. *Id.* The participation in communication is not to be hindered by compulsion. *Id.*

131. *See* Solum, *supra* note 121, at 262-63 (highlighting that treating people with dignity and respect is important in decision-making process).

132. *Id.* at 264 (combining dignity, autonomy, and equality for the notion of participation).

133. This is the concept of *motive-based trust. See, e.g.*, Tyler, *supra* note 104, at 294.

134. Tyler, *Procedural Justice and the Courts*, *supra* note 121, at 31.

135. Hough et al., *supra* note 121; Jackson et al., *supra* note 121; Mazerolle et al., *supra* note 121; Tyler & Wakslak, *supra* note 121; Sunshine & Tyler, *supra* note 125; Meares et al., *supra* note 121; Tyler, *supra* note 104, at 286.

136. Tyler, *supra* note 104, at 324-27; Meares et al., *supra* note 121, at 314-20.

137. Tyler, *supra* note 104, at 326.

inferences about the motives of police behavior.[138] If people experienced fairness from the police or generally consider the police to act fair, they are less likely to infer that profiling occurs.[139] Quality of decision-making, quality of treatment, and inferences of trustworthiness were found to affect whether people inferred that they have been profiled.[140] If the police treats people in a respectful, polite, and fair manner, they will be trusted. However, if police officers are perceived as disrespectful and impolite, people will distrust their motives and are more likely to infer illegitimate motives like racism. The perception of fairness, and consequently the perceived legitimacy, do not depend on the legality of police behavior.[141] It is therefore not necessarily relevant for the perception of fairness that the police is legally authorized to take an action or decision. In the context of proactive policing, studies suggest that proactive police contact does not necessarily deter from criminal behavior. It can however make members of heavily policed areas feel like subjects of suspicion which will impair the view of police legitimacy.[142] The high police presence might be perceived as a signal of disrespect and distrust, which will lower people's willingness to cooperate and communicate with legal authorities. As with racial profiling, the people will distrust the motives guiding police action. Consequently, proactive policing can undermine the relationship between people and police. The same holds true for predictive policing, especially when the software sends police officers repeatedly and more frequently to neighborhoods with a high amount of ethnic or racial minorities.[143] By and large, the theory of procedural justice may offer valuable insights on the acceptance of legal decisions that could be harnessed to enhance the acceptance of ADM.

C. THE RULE OF LAW AND DUE PROCESS

The rule of law and the principle of due process provide further criteria for assessing automated legal decision-making. Both concepts are rather vague, comprising a multitude of aspects, and are constantly evolving. The core notion of the rule of law is that the law governs individual and institutional behavior, even the behavior of legal authorities and legislators.[144] Due process, a subcategory of the rule of law, concentrates on

138. Tyler & Wakslak, *supra* note 121.

139. *Id.* at 276.

140. *Id.* at 277.

141. Meares et al., *supra* note 121, at 318-20.

142. Tom R. Tyler et al., *The Consequences of Being an Object of Suspicion: Potential Pitfalls of Proactive Police Contact*, 12 J. EMPIR. LEG. STUD. 602 (2015).

143. A *feedback loop* is created, meaning that with every arrest in a neighbourhood, the software will estimate a higher chance of more crimes in that area. In other words, the crime detection rate is raised disproportionately to the actual crime rates. Matt Reynolds, *Biased Policing Is Made Worse by Errors in Pre-Crime Algorithms*, NEW SCIENTIST (Oct. 4, 2017), https://www.newscientist.com/article/mg23631 464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/ [https://perma.cc/KB69-2 DAK].

144. *Cf. Rule of law*, DICTIONARY BY MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/rule%20of%20law [https://perma.cc/S2LB-MXMD]; *Rule of law*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/topic/rule-of-law [https://perma.cc/CN7U-4HAF]; Ted Honderich, *Rule of Law*, THE OXFORD COMPANION TO PHILOSOPHY 824 (2005).

legal proceedings, the adherence to the relevant rules governing the proceedings, and the protection of individual rights during the proceedings.[145] The most crucial aspect of due process is the right to be heard, giving every party the right to present their arguments before court.[146] Another important aspect is the right to notice of an action or decision to be taken, as well as justifying reasons.[147] Giving reasons for a decision can reveal flaws or errors in the decision-making process, for example, if the legal authority was acting unfairly or outside its competence, and provide information for challenging it.[148] A fair trial also means equality of arms: every party shall have the opportunity to present their case, under conditions that do not place them at a disadvantage vis-à-vis their opponent.[149] The rule of law, or due process, are multifaceted, and many more aspects can be subsumed under these principles. As of late, issues pertaining to privacy are also discussed in the context of due process.[150] For example, the General Data Protection Regulation ("GDPR") adopted by the European Union provides a right of access[151] by the data subject:

> "The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: […] the existence of automated

145. *Due process*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/topic/due-process [https://perma.cc/2EWJ-AWC7]. The term is used in a broader sense than the Due Process Clause under the United States Constitution.

146. The Right to be Heard is codified in Art. 6 of the European Convention on Human Rights. European Convention for the Protection of Human Rights and Fundamental Freedoms, art. 6, Nov. 4, 1950, 213 U.N.T.S. 222 [hereinafter European Convention on Human Rights] (entered into force Sept. 3, 1953). In some legal systems, a right to be heard is also granted in administrative law (e.g., concerning certain police action), such as in Germany's § 28 of the Administrative Procedure Act (both federal and state law). Concerning the United Kingdom, *cf.* Marion Oswald, *Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power*, 376 PHIL. TRANS. SERIES A 1, 5 (2018). According to Henry J. Friendly, a hearing comprises of 11 elements: (1) an unbiased tribunal; (2) notice of the proposed action and the grounds asserted for it; (3) an opportunity to present reasons why the proposed action should not be taken; (4) the right to call witnesses; (5) the right to know the evidence against oneself; (6) the right to have the decision based only on the evidence presented; (7) the right to counsel; (8) the making of a record; (9) a statement of reasons; (10) public attendance; and (11) judicial review. Henry J. Friendly, *Some Kind of Hearing*, 123 U. PA. L. REV. 1267, 1279-95 (1975).

147. *Cf.* Friendly, *supra* note 146, at 1280-81.

148. Mireille Hildebrandt, *Privacy As Protection of the Incomputable Self: Agonistic Machine Learning*, 20 THEOR. INQ. LAW 83, 119 (2019); Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1120-26 (2018); Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for*, 16 DUKE L. & TECH. REV. 18, 54 (2017).

149. European Convention on Human Rights, art. 6(1); Öcalan v. Turkey, 2005-IV Eur. Ct. H.R. 133, 179.

150. Citron, *supra* note 5. Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process For Automated Predictions*, 89 WASH. L. REV. 1, 18-30 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014). *Cf.* Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"*, AI MAGAZINE, Oct. 2, 2017, at 50 (the academic discussion of the General Data Protection Regulation (GDPR)); Edwards & Veale, *supra* note 148 (discussing whether Art. 15 of the GDPR is an appropriate tool to access information on certain decisions); Edwards & Veale, *supra* note 148 at 65-74 (presenting alternative remedies in Art. 17 and Art. 20 of the GDPR). Stefanie Hänold, *Profiling and Automated Decision-Making: Legal Implications and Shortcomings*, ROBOTICS, AI AND THE FUTURE OF LAW 123, 132-45, 147-50 (Marcelo Corrales et al. eds., 2018).

151. Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017).

> decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[152]

Moreover, the Right to Erasure under Article 17 of the GDPR enables the individual data subject to have personal data erased, withdrawn from a model (dubbed "Machine Unlearning"), or have the model itself erased if it represents the data subject's personal data.[153] In addition, the GDPR accords the "right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" to data subjects.[154] However, Article 22 of the GDPR is not applicable if decision-support tools are used by a human in the loop.[155]

D. MERGING THE THREE FRAMEWORKS

The three frameworks discussed—machine ethics, procedural justice, and the rule of law and due process—address different aspects of interaction between citizens and legal authorities. Machine ethics introduce normative criteria, whereas the theory of procedural justice tries to determine criteria based on empirical investigation. While machine ethics and the (socio-psychological) theory of procedural justice constitute interdisciplinary approaches, the rule of law and due process represent constitutional principles.

Some of the aforementioned criteria overlap, for example *neutrality*, *impartiality*, and *fairness* can be viewed together as they all deal with issues of discrimination or bias. The *right to be heard* corresponds to Tyler's notion of *voice*. When decisions or actions are proposed by ADM-software, the viewpoint and the arguments of the affected citizens might not receive adequate consideration by legal authorities. ADM-software does not automatically provide an opportunity for citizens to state their case. *Trust* and *accountability* both require legal authorities to responsibly deploy ADM-software. The use of ADM-software should not cause harm to humans. Extraneous or intolerable motives or factors should have no impact on the decision. Citizens need to be able to trust in the sincerity of the decision-making process, or at least have the means to challenge flawed ADM-software. *Respect* and the concept of *human dignity* can also be combined to assess whether citizens might feel objectified by the authorities' decision-making. Individuals should not be deprived of their autonomy and self-

152. Regulation 2016/679 of the European Parliament and of the Council of Apr. 27, 2016 on the Protection of Nat-ural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 15, para. 1, 2016 O.J. (L 119) 33 (EU) [hereinafter GDPR].

153. Edwards & Veale, *supra* note 148, at 71-72 (discussing the possibility to have the data or model erased if it has been traded, which might be of less relevance in the public sector).

154. GDPR, art. 22, para. 1.

155. Hänold, *supra* note 150, at 133-32, 147.

determination. *Transparency* and *explicability* require that individuals can obtain a meaningful explanation of an algorithmic decision concerning them.

Some of these criteria are interdependent: Citizens will only feel that they are treated with respect and dignity, if they feel treated fairly and have the chance to state their case. Discrimination or bias in the decision-making process lead to a loss of trust. Transparency is a fundamental prerequisite for evaluating the fairness of ADM-software, and for attributing accountability. Furthermore, citizens can only effectively challenge a decision and present arguments in their favor, if they know the decision's rationale.

IV. FORESTALLING THE KAFKAESQUE

To prevent the evolution of algorithmic decision-making into a Kafkaesque bureaucracy or a Kafkaesque justice system, safeguards need to be established to ensure the fairness, accountability, and transparency of the decision-making process. The citizens affected must have a right to be heard and should be treated with dignity and respect.

A. TRANSPARENCY

Due to the abstract and complex structure of ML algorithms, they lack transparency.[156] Yet, explaining their function and their results is (within a limited scope) feasible (as opposed to human motivations and rationales, that are even more opaque and not easily accessible).

1. The Opacity of Machine Leaning Algorithms

ML algorithms are perceived as "black boxes".[157] Their opacity arises from the difficulty to trace the process of turning input into output. The determinants leading to an outcome can hardly be extracted since the reasoning process is not represented by ML algorithms, but influenced by its many layers and nodes, or by multiple decision trees composing the Random Forests result. This form of opacity can be called *opacity as the complexity of scale*.[158] The intricate design of ML algorithms, their complex learning technique, and the large amounts of data influencing the algorithm's structure, render it difficult to trace why a specific result was obtained.[159] Other circumstances additionally contribute to the perception of ML algorithms as "black boxes". Opacity is also intentionally caused to protect

156. ALPAYDIN, *supra* note 22, at 155-56; Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1, 5-6, 9 (2016) ("[O]pacity as the complexity of scale" is the third of three forms of opacity compiled by Burrell); Szymielewicz et al., *supra* note 17.

157. This problem is well discussed. *E.g.*, Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1 (forthcoming 2019), https://papers.ssrn.com/sol3/papers.cf m?abstract_id=3293008## [https://perma.cc/TPT6-PARP]; Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1167, 1205-13 (2017); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 6-7 (2017); Oswald, *supra* note 146, at 3-4; Lehr & Ohm, *supra* note 22, at 705-10; Szymielewicz et al., *supra* note 17.

158. *Cf.* Burrell, *supra* note 156, at 5-6, 9 ("[O]pacity as the complexity of scale" is the third of three forms of opacity compiled by Burrell).

159. Burrell, *supra* note 158, at 5-6; Zalnieriute et al., *supra* note 56, at 15-16.

trade secrets of companies or state secrets.[160] The algorithms could be elucidated by either using *open-source* software instead of proprietary software,[161] or by enhancing regulation and audit.[162] Furthermore, algorithms can appear opaque to the general public that is not familiar with the design of algorithms or their underlying code.[163] Disclosing information about a system will avail them nothing, unless they seek expert advice to extract the relevant information.[164] The opacity of ML algorithms can lead to the perception of ADM-systems as faceless, anonymous, sinister entities, impairing the societal acceptance of their decisions. But are they more opaque than human decision-makers?

2. Human Decision-Making as a Benchmark for Transparency?

Retracing algorithmic decisions, though complex and scientific, is possible, contrary to disclosing the factors that lead to a human decision.[165] The ideal of judges merely applying the law by using formalistic logic has been abandoned for about a century now.[166] Judicial decision-making is influenced by attitudes, preferences, and political views.[167] The outcome of judicial decisions may also depend on the effort put into the reasoning process and the perception of accountability towards the public, colleagues, the parties involved, and the profession.[168] Other factors, like emotions, unconscious assumptions, fatigue, etc., influence legal decisions as well, albeit judges tend to underestimate their impact.[169] Legal decision-makers suffer from the same cognitive failings as humans in general. Of special significance is the proneness to the *anchoring effect*:[170] the initial suggestion rendered by the ADM will cause the decision-maker to unconsciously align his decision with the result of the ADM-support tool and will thereby influence the decision-

160. Burrell, *supra* note 158, at 3-4; Zalnieriute et al., *supra* note 56, at 14.

161. Zalnieriute et al., *supra* note 56, at 14-15. This was discussed in the context of COMPAS, as the company (Equivant) was not obliged to disclose its algorithm. State v. Loomis, 2016 WI 68, 371 Wis. 2d 235, 881 N.W.2d 749.

162. Burrell, *supra* note 158, at 4; Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 109-10 (2017). A third-party audit might however not achieve more transparency than internal validation by the administrative agency deploying the ADM-software.

163. Burrell, *supra* note 158, at 4 (referencing "[o]pacity as [t]echnical [i]lliteracy").

164. Zalnieriute et al., *supra* note 56, at 15.

165. Kleinberg et al., *supra* note 7, at 10-13. Whether judges' reasoning processes are equivalent to lay decision-makers has been the subject of debate. *Cf.* Frederick Schauer, *Is There a Psychology of Judging?*, *in* THE PSYCHOLOGY OF JUDICIAL DECISION MAKING 103, 103 (David E. Klein & Gregory Mitchell eds., 2010).

166. *Cf.* Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457 (1897); Lawrence Baum, *Motivation and Judicial Behavior: Expanding the Scope of Inquiry*, *in* THE PSYCHOLOGY OF JUDICIAL DECISION MAKING 3-4 (David E. Klein & Gregory Mitchell eds., 2010).

167. With regard to the Supreme Court of the United States: *see, e.g.*, Jeffrey A. Segal & Albert D. Cover, *Ideological Values and the Votes of U.S. Supreme Court Justices*, 83 AM. POL. SCI. REV. 557 (1989).

168. Brandon L. Bartels, *Top-Down and Bottom-Up Models of Judicial Reasoning*, *in* THE PSYCHOLOGY OF JUDICIAL DECISION MAKING 41, 48 (David E. Klein & Gregory Mitchell eds., 2010). *Cf.* Baum, *supra* note 166, at 15-16.

169. Sourdin, *supra* note 55, at 1128-29; Daniel L. Chen, *Machine Learning and the Rule of Law*, *in* LAW AS DATA 433, 433-34 (Michael A. Livermore & Daniel Rockmore eds., 2019) (listing the relevant research).

170. Daniel Kahneman, THINKING, FAST AND SLOW 119-28 (2012); Schauer, *supra* note 165, at 113.

maker's judgment. A judge may justify his decision *ex post facto*, while the reasons for reaching the decision remain undisclosed.[171] Another question is the level of scrutiny applied in the decision-making process. While a systematic and analytical processing of facts, evidence, and the relevant law might seem to be the obvious mode of reasoning, decision-makers will often use heuristics to deal with high workload or simple and repetitive cases.[172] Legal reasoning, in a perfect world, would entail the analysis of information with objective scrutiny and basing the decision solely on facts and evidence – which would be bottom-up or inductive reasoning.[173] Top-down or deductive reasoning is dictated by predispositions and therefore tends to be biased.[174] Human and judicial decisions are rarely just deliberative and systematic, or only spontaneous and impetuous.[175] Often, decisions are rendered in hybrid processes in which heuristic and systematic, top-down and bottom-up reasoning coincide,[176] or in a mixed controlled process in which potentially biasing influences are partly overcome by deliberation.[177] Objective reasoning can be induced if the decision-maker is afraid that her decision will be considered invalid. The feeling of accountability, or a propensity for self-presentation, can motivate the decision-maker to improve the accuracy of her decision. However, a strong and convincing justification of a decision is no evidence for an objective decision.[178] The "*illusion of objectivity*"[179] is cast upon justifications that are constructed *post hoc*, after the desired outcome has already been determined.

But do we ask more of algorithms than we expect of human decision-makers? Human decisions might be inscrutable, but they are not unexplainable. Professionals normally work within a regulated structure and are subject to oversight over their actions and decisions. [180] They can be questioned and rebuked in case of unlawful behavior.[181] Marion Oswald proposes to not adopt a higher standard for algorithms than for humans, but one adapted to the decision-rendering process of ADM.[182] Determining an appropriate standard of transparency for ADM depends in part on the technical possibilities and limitations on explaining results of ML algorithms.

171. Zalnieriute et al., *supra* note 56, at 13; Richard E. Nisbett & Timothy D. Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231 (1977).

172. Baum, *supra* note 166, at 17.

173. Bartels, *supra* note 168, at 44.

174. *Id.* at 44.

175. *Id.* at 48.

176. *Id.*

177. *Id.* at 44-45.

178. *Id.* at 46.

179. *Id.*; Ziva Kunda, *The Case for Motivated Reasoning*, 108 PSYCHOL. BULL. 480, 482-83 (1990); *cf.* Schauer, *supra* note 165, at 110.

180. Oswald, *supra* note 146, at 6.

181. Frank Pasquale & Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviourism*, 68 U. TORONTO L. J. 63, 66 (2018); Oswald, *supra* note 146, at 6.

182. Oswald, *supra* note 146, at 6-7.

3. xAI (Explainable AI)

A whole field of research is concerned with "opening the black box" of ML algorithms.[183] Correlation between different features can be detected relatively easily by changing the respective information and comparing the outputs.[184] Key information about an ML system, especially the training data sets and the techniques used, can be disclosed.[185] Alternatively, the performance of a system can be measured by statistically evaluating the accuracy of their outcomes.[186] Explainability of ML algorithms can be enhanced by limiting their complexity.[187] For example, by reducing input variables, or by choosing a model that can be interpreted more easily, such as opting for Decision Trees over ANN.[188] However, some tasks require many input variables and complex structures (e.g. image recognition). In such cases, the input variables do not correspond to a concept that humans understand which makes any explanation of the result very difficult.[189]

Most approaches of explaining ML algorithms are only meaningful to experts and not practical for non-experts like citizens or legal professionals.[190] The objective of many xAI-researchers is to render ML algorithms explainable or interpretable in terms understandable for humans.[191] *Local approximations*[192] of decisions create an approximation of a decision-making algorithm, which receives the same input and models the decision.[193] One can explain a specific decision (rather than the systems overall behavior), by

183. The literature on this subject is vast. Good summaries on the Explainable AI-research (xAI) are: Mittelstadt et al., *supra* note 51, at 280-83.

184. Doshi-Velez et al., *supra* note 107; *see id.* at 3 ("By looking at the effect of changing that information on the output and comparing it to our expectations, we can infer whether it was used correctly").

185. Zalnieriute et al., *supra* note 56, at 14.

186. Doshi-Velez et al., *supra* note 107, at 10-11. Another method to hold AI accountable is providing theoretical guarantees that govern the system's processes. However, real-world settings can rarely be formalized to warrant the proper functioning of such a system. The advantage of ML algorithms is that they perform well where traditional, strictly rule-based algorithms do not.

187. Lehr & Ohm, *supra* note 22, at 692-93; Selbst & Barocas, *supra* note 148, at 1110-15.

188. Selbst & Barocas, *supra* note 148, at 1110-13; Veale et al., *supra* note 64, at 4. A target percentage of accuracy (e.g., 75%) can be established to avoid trading off transparency for accuracy.

189. Edwards & Veale, *supra* note 148, at 59-61. Been Kim et al., *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), in* INT'L CONF. MACHINE LEARNING 2018 1 (2018), https://arxiv.org/abs/1711.11279 [https://perma.cc/2FQX-53FT] (devising a method to quantify the degree to which a user-defined concept is important to a classification by basically training another algorithm (so-called *Concept Activation Vector*)). *See also* Amirata Ghorbani et al., *Automating Interpretability: Discovering and Testing Visual Concepts Learned by Neural Networks* (Feb. 7, 2019), https://arxiv.org/abs/1902.03129 [https://perma.cc/9SM9-W8LL].

190. Mittelstadt et al., *supra* note 51, at 279, 285-86.

191. Finale Doshi-Velez & Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, 2 (2018), https://arxiv.org/abs/1702.08608 [https://perma.cc/5SGB-SMEK].

192. Doshi-Velez et al., *supra* note 107, at 7; *see, e.g.*, Marco Tulio Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* (Feb. 16, 2016), https://arxiv.org/ abs/1602.04938 (proposing *Local Interpretable Model-Agnostic Explanations* (meaning that the predictions of any model)). Local approximations are subject to limit: a trade-off between the understandability of the function, the quality of the approximation, and the size of the domain for which the approximation applies, *cf.* Sandra Wachter et al., *Counterfactual Explanations Without Opening The Black Box: Automated Decisions and the GDPR*, 31 HARV. J. L. & TECH. 841, 851 (2018). If lay people are supposed to make use of local explanations, they need to be educated about the limitations and the unreliability of the rendered explanations. *Id.*

193. Wachter et al., *supra* note 192, at 850-51. "As multiple outcomes based on changes to multiple variables may be possible, a diverse set of counterfactual explanations should be provided, corresponding to different choices of nearby possible worlds for which the counterfactual holds." *Id.* at 881.

probing the inputs to a system and determining, which factors have the greatest effect on the outcome. For different instances (e.g. for different defendants applying for parole), different factors influence the decision. In contrast, *global* models[194] investigate the "algorithm-wide" importance of certain input variables across many decisions and queries.[195]

Contrastive or *counterfactual explanations* can disclose whether a certain factor influenced or determined the outcome, and which factors would have led to a different outcome.[196] For example, as counterfactual explanation of a recidivism score could inform the defendant that his history of non-compliance during imprisonment decisively influenced his high risk score. If he wants to lower his risk score, he might display more compliant behavior. Counterfactual explanations meet our needs for causal explanations and allow for information exchange and the discussion of the result's justifiability.[197]

Counterfactual explanations can be given by an *explanation system* that is distinct from the ADM-system.[198] The explanation system receives the same input as the ADM-system, and outputs its own result, trying to predict the same result as the ADM-system.[199] Both the result from the ADM-system and the explanation system can be subsequently compared. Counterfactual explanations provide only a minimal amount of information which might not suffice to adequately address the individual's need for an explanation.[200] It is not helpful to obtain information about an immutable factor that influenced the decision (like a high recidivism score because of the defendant's sex), especially if other factors contributing to the decision are alterable (such as the compliance during imprisonment). As a corrective, multiple diverse counterfactual explanations can be provided to the person affected by the decision.[201] Counterfactual explanations enable individuals to obtain an explanation that can help them understand a decision, adjust their behavior, and possibly achieve the desired decision.[202] Moreover, the counterfactual explanation might report back that protected variables, like race or sex, affected the decision, so discriminatory tendencies in the algorithm can be detected.[203] Counterfactuals can either be generated automatically at the

194. Edwards & Veale, *supra* note 148, at 55-56 (referring to such explanations as "*model-centric*").

195. Lehr & Ohm, *supra* note 22, at 708-09; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 708-09 (2017). The output of a global model is a *variable importance plot* displaying the relative importance of different input variables.

196. Mittelstadt et al., *supra* note 51, at 283-86.

197. Doshi-Velez et al., *supra* note 107, at 7; Mittelstadt et al., *supra* note 51, at 283-86; Wachter et al., *supra* note 192, 851-52; *cf.* Matt J. Kusner et al., *Counterfactual Fairness* (Mar. 20, 2017), https://arxiv.org/abs/1703.06856 [https://perma.cc/7CJD-68PE].

198. Doshi-Velez et al., *supra* note 107, at 7-8.

199. *Id.* at 7.

200. Wachter et al., *supra* note 192, at 851.

201. *Id.* at 851. "As multiple outcomes based on changes to multiple variables may be possible, a diverse set of counterfactual explanations should be provided, corresponding to different choices of nearby possible worlds for which the counterfactual holds." *Id.* at 881.

202. Wachter et al., *supra* note 192, at 844.

203. *Id.* at 853.

time a decision is rendered, or, if an explanation is requested, later based on an archived copy of the model.[204] Sandra Wachter et al. proposed an automated implementation of counterfactuals, so that explanations are calculated and disclosed in an automated way.[205] Counterfactual explanations enable the explanation of decision without infringing on the privacy of individuals whose data is contained in the training dataset, or on trade secrets.[206] They admittedly do not suffice if the functionality of the system, or the rationale of a decision needs to be revealed, since no statistical evidence to assess the algorithms with respect to biases is provided.[207] It should be noted that ADM-systems have to be designed to store inputs, intermediate steps, and outputs.[208] Alternatively, a *pedagogical* or *interactive* system can allow for the exploration of a ML system with human-computer interaction.[209] Citizens could test a scoring system and see what happens if they change certain features.[210]

4. Explaining Legal Decisions

Transparency of legal decision-making benefits both citizens and decision-makers. Explanations can enable legal authorities to understand how decision-support systems obtain their results.[211] If decision-makers can interpret the outcome of ADM-software, they can evaluate whether the recommended decision should be implemented, double-checked, or dismissed due to potential flaws.[212] For example, a police officer should be able to assess whether an action recommended by ADM-software is in accordance with the law.[213] The reasons for a legal decision must be expounded intelligibly to the affected individuals to foster their confidence in the fairness of the decision-making process, and to enable them to challenge erroneous decisions.[214]

A legally meaningful explanation must contain several aspects.[215] The main factors leading to a decision should be revealed, but it should also be disclosed whether changing a certain factor would have led to a different

204. *Id.* at 881.

205. *Id.* (proposing "auditing APIs" allowing users to request counterfactual explanations from the service provider).

206. *Id.* at 882-83.

207. *Id.* at 883.

208. Contrary to humans who generally store information needed to explain their decisions. *See* Doshi-Velez et al., *supra* note 107, at 9-10.

209. Edwards &Veale, *supra* note 148, at 61-64 (discussing the danger of people trying to "game" the system); *see also* Selbst & Barocas, *supra* note 148, at 1115-17.

210. *Cf.* Citron, *supra* note 5, at 28-29 (*Interactive Modelling*).

211. Decision-makers need to "make sense" of the ADM-model. *See* Max van Kleek et al., *The Need for Sensemaking in Networked Privacy and Algorithmic Responsibility*, SENSEMAKING WORKSHOP CHI 2018 1, 6-7, https://drive.google.com/file/d/1-FWI4BnLfxnRuM40KIgu31KDbrCy8Z-k/view [https://perma.cc/LX6M-54EE].

212. Oswald, *supra* note 146, at 8. To be able to do so, decision-makers also need to understand measures of accuracy and performance metrics. *See* Veale et al., *supra* note 64, at 9-10.

213. Oswald, *supra* note 146, at 7.

214. Doshi-Velez & Kim, *supra* note 191, at 2; Oswald, *supra* note 146, at 7.

215. Doshi-Velez et al., *supra* note 107, at 3. C*f.* the concept of Counterfactual Explanations: Wachter et al., *supra* note 192, at 843; Mittelstadt et al., *supra* note 51, at 283-84.

decision. Furthermore, it may be important to know why similar-looking cases might be decided differently to assess the consistency of the decision-making process.[216] The demands concerning the reliability and scope of the explanation can be determined by balancing the necessity of an immediate decision and the magnitude of its impacts, considering both practicability and the rights of the individuals affected.[217] It is not always necessary to expose the internal logic, the training data, or the source code of an ADM-system in order to build trust and achieve societal acceptance of ADM.[218]

Explanations are not a universal remedy to legitimize ADM, because they might cast the "*illusion of clarity*" when a genuine justification of an algorithmic decision cannot be obtained.[219] Even if ADM is relatively transparent, people are not always capable to make use of the explanation given to them. Preaching transparency as panacea mean succumbing to the "*transparency fallacy*".[220] The individuals affected by an algorithmic decision lack the time, resources, and the required expertise to exercise their rights.[221] The main focus should be on designing ADM as carefully and reliably as possible, since most affected individuals do not want an explanation, but rather wish for the decision or action to not have occurred.[222] Transparency can be fostered by deciding against complex methods like ANNs in favor of decision trees, or by including less variables.[223]

B. THE QUEST FOR ALGORITHMIC FAIRNESS

As already mentioned, *algorithmic biases* have been vigorously discussed in the context of facial recognition and risk assessment software.[224] In the context of the discussion on the recidivism scoring software COMPAS, some light will be shed on algorithmic biases and how to ensure algorithmic fairness.

1. Biased Algorithms?

When ProPublica analyzed COMPAS, it found that black defendants were not only more likely to be classified as likely to reoffend, but they were

216. Doshi-Velez et al., *supra* note 107, at 3.

217. Oswald, *supra* note 146, at 8-9.

218. Wachter et al., *supra* note 192, at 843. Edwards & Veale, *supra* note 148, at 64-65 (calling this a "*decompositional* explanation," revealing the inner structures, such as the weights, neurons, decision trees and architecture).

219. Desai & Kroll, *supra* note 157, at 4-5 (explaining why a face was matched with that of a wanted person will be difficult).

220. Edwards & Veale, *supra* note 148, at 67.

221. "Individuals are mostly too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of these individual rights." *Id.*

222. Edwards & Veale, *supra* note 148, at 42.

223. Veale et al., *supra* note 64, at 4 (a target percentage of accuracy (*e.g.*, 75%) can be established to avoid trading off transparency for accuracy).

224. For another example of algorithmic bias, *see* Isobel Asher Hamilton, *Amazon Built an AI Tool to Hire People But Had to Shut It Down Because It Was Discriminating Against Women*, BUSINESS INSIDER (Oct. 10, 2018), https://www.businessinsider.de/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10?r=US&IR=T [https://perma.cc/4UN4-7ECV]. *Cf.* Crawford& Schultz, *supra* note 150, at 99-101, 103-05.

misclassified as having a higher risk to reoffend (twice as often as white defendants), whereas white defendants were misclassified as low risk almost twice as often as black offenders.[225] Subsequently, the use of software like COMPAS was subject of criticism. Firstly, because some persons classified as black are more likely to be re-arrested does not justify predicting other persons (especially from a different city or state and acting in a different period of time) to be classified as likely to re-offend.[226] The algorithms predicting recidivism or calculating the recidivism score are trained with historical data meaning social issues are cemented rather than overcome or reduced.[227] Second, correlations and causations get mixed up:[228] a certain ethnicity is not the reason for higher crime and recidivism rates. They rather correlate with a series of other issues such as poverty, lack of opportunities, and living in poor districts.[229] Even if an algorithm is formally blind to race or sex it might be using a correlated proxy.[230] However, the results of COMPAS can hardly be challenged. The algorithm generating the recidivism score is opaque, partly due to the methods used in these algorithms, working on such high levels of abstraction that the factors leading to the decision and the extent of their influence on the decision can hardly be revealed. Additionally, these algorithms are often private intellectual property and are not disclosed.[231] Lastly, the performance of the COMPAS algorithm was criticized as suboptimal: some studies suggest that humans[232] or simpler algorithms[233] taking into account only few factors when assessing recidivism outperform COMPAS, or at least achieve similar results.[234]

225. Angwin et al., *How We Analyzed the COMPAS Recidivism Algorithm*, *supra* note 92.

226. Aaron M. Bornstein, *Are Algorithms Building the New Infrastructure of Racism?*, NAUTILUS (Dec. 21, 2017), http://nautil.us/issue/55/trust/are-algorithms-building-the-new-infrastructure-of-racism [https://perma.cc/J9EU-DZK6].

227. Hao, *supra* note 90.

228. *See* Kleinberg et al., *supra* note 7, at 20 (ML algorithms cannot infer causalities).

229. Hao, *supra* note 90.

230. Doshi-Velez et al., *supra* note 107, at 8; Kleinberg et al., *supra* note 7, at 20.

231. Kleinberg et al., *supra* note 7, at 32; Rebecca Wexler, *When a Computer Program Keeps You in Jail*, THE NEW YORK TIMES (June 13, 2017), https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html?module=inline [https://perma.cc/A2NH-C9RY].

232. Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*. 4 SCI. ADV. eaao5580 (2018). Dressel and Farid conclude that two features used to predict recidivism achieve the same level of accuracy as the 137 features taken into account by COMPAS. *Id.* They reviewed nine algorithmic approaches and found that only moderate levels of accuracy could be achieved, challenging the use of predictive algorithms in criminal justice decision-making. *Id.*; *see also* Sarah Tan et al., *Investigating Human + Machine Complementarity for Recidivism Predictions* 6 (Aug. 28, 2018), https://arxiv.org/abs/1808.09123 [https://perma.cc/N9YC-38LB] (examining whether humans and COMPAS make similar assessments concerning the same defendants or whether human reasoning differed from algorithmic results. They conclude: "Ultimately, we want to leverage the best of both worlds: humans that glean subtle, interpersonal insights from rich context, and machine algorithms that provide rigor and consistency. humans that glean subtle, interpersonal insights from rich context, and machine algorithms that provide rigor and consistency.").

233. *Cf.* Elaine Angelino et al., *Learning Certifiably Optimal Rule Lists for Categorical Data* (Apr. 6, 2017), https://arxiv.org/abs/1704.01701 [https://perma.cc/UCZ7-BUCC]; Jongbin Jung et al., *Simple Rules for Complex Decisions* (Feb. 15, 2017), https://arxiv.org/abs/1702.04690 [https:// perma.cc/C96T-L3JC] (whose model outperformed judges); Jiaming Zeng et al., *Interpretable Classification Models for Recidivism Prediction* (Mar. 26, 2015), https://arxiv.org/abs/1503.07810 [https:// perma.cc/XK6J-E2FT].

234. Ed Yong, *A Popular Algorithm Is No Better at Predicting Crimes Than Random People*, THE ATLANTIC (Jan. 17, 2018), https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-

However, in more complex scenarios with access to extensive information and no immediate feedback, algorithmic risk assessment instruments were found to perform better in a recent experiment conducted by Zhiyuan Lin, et al.[235]

The criticism of COMPAS ultimately led to a general discussion of algorithmic bias and algorithmic fairness.[236] While the consideration of criteria like "race" or ethnic affiliation was widely condemned, others raised the question whether statistically relevant factors should be taken into account if accuracy can be improved without trying to figure out as to why the result was obtained.[237] Not including it would lead to a distortion and misrepresentation of data and less accurate or even arbitrary results.[238] Yet, protected criteria like race and ethnicity are naturally predetermined factors.[239] The defendant has no influence, so basing a far-reaching decision on such criteria is unfair. Furthermore, considering race as a causal factor for higher recidivism is unjustified: studies suggest that other socioeconomic factors such as poverty, discrimination, and high income inequality increase the likelihood of committing crimes and are typically affecting ethnic minorities)[240] Interestingly, Jon Kleinberg et al. have voiced the idea of using the discriminatory criterion ("race") to mitigate the racial bias in the data.[241]

When adjusting ML algorithms, the following trade-off has to be made: either one reaches the same overall accuracy for all groups, meaning members of one group (black defendants) are more likely to be misclassified (as likely to reoffend), or one calibrates the algorithm to lower the rate of misclassfication of black defendends in exchange for a lower overall

algorithm/550646/ [https://perma.cc/VR5N-956E] (raising the question of whether predicting recidivism will ever get significantly better than a coin toss).

235. Zhiyuan Lin et al., *The Limits of Human Predictions of Recidivism*, 6 SCIENCE ADVANCES 1 (2020); *cf.* Sophie Bushwick, *Will Past Criminals Reoffend? Humans Are Terrible at Guessing, and Computers Aren't Much Better*, SCIENTIFIC AMERICAN (Feb. 14, 2020), https://www.scientificamerican. com/article/will-past-criminals-reoffend-humans-are-terrible-at-guessing-and-computers-arent-much-better/ [https://perma.cc/YAN8-TRLV].

236. *See, e.g.*, Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 2 (Sep. 19, 2016), https://arxiv.org/abs/1609.05807 [https://perma.cc/KDU6-2CSN].

237. Richard A. Berk & Justin Bleich, *Statistical Procedures for Forecasting Criminal Behaviour*, 12 CRIMINOL. PUB. POL'Y 513, 516-17 (2013).

238. If the district of residence is still a factor influencing the recidivism score, and the district is mainly inhabited by an ethnic group with higher crime and recidivism rates, a person living there might be rated to be more likely to reoffend, despite not belonging to the ethnic minority. This person would be treated worse based on the recidivism rates of his district although he does not actually belong to the group of people with a higher risk to reoffend. The predictions would in this case go awry. The decision-makers could alternatively resort to other forms of information, if one type of information is banned. *Cf.* Kleinberg et al., *supra* note 7, at 28.

239. Kleinberg et al., *supra* note 7, at 8-9.

240. *E.g.*, Robert Agnew, *A General Strain Theory of Community Differences in Crime Rates*, 36 J. RES. CRIME DELINQ. 123, 133-34 (1999) (discussing the effect of experienced racial discrimination on criminal behavior). Michael J. Hindelang, *Variations in Sex-Race-Age-Specific Incidence Rates of Offending*, 46 AM. SOCIOL. REV. 461 (1981) (stating that sex, race, and age are relevant variables for indicating offensive behavior. Hindelang nevertheless asserts that people of ethnic minorities may have differential (meaning fewer) opportunities for jobs and increasing income and status in comparison to white Americans).

241. Kleinberg et al., *supra* note 7, at 34. For example, the recidivism score can be adjusted for black defendants by lowering the weight of the neighborhood and increasing the weight of more significant factors such as compliance during imprisonment. The importance of residence is generally problematic since it can lead to social discrimination. Veale et al., *supra* note 64, at 6-7.

accuracy.[242] COMPAS might be fair in the sense that risk scores have the same overall accuracy for black and white defendants (*predictive parity*).[243] But the rate of misclassifications was not the same for these different groups (*error rate balance*).[244] Essentially, the software-provider Equivant approaches the assessment of recidivism from the perspective of reoffending defendants, trying to minimize false negatives, whereas ProPublica conveys the perspective of defendants not reoffending, attempting to avoid false positives.[245] The fundamental question is therefore whether we want to reduce false negatives and enhance public safety or reduce false positives and prevent defendants from being wrongfully denied parole.[246]

The bias in algorithms can be caused by many different flaws at different stages of the algorithm's development.[247] The selection of the outcome measure or the input variables can introduce discrimination if they are typically fulfilled by one group, but not by another. Too little information can distort the results (sampling bias).[248] Biases can be contained in the training data, especially when it comprises past human decisions reflecting prejudice or implicit bias.[249]

2. Enhancing Algorithmic Fairness

Fairness and neutrality of algorithmic decisions can be achieved if the software is prudently and conscientiously designed and deployed. Carefully designed software will display less bias than human alternatives, because the bias of algorithmic decisions does not originate from the algorithm itself but from its design, the training data, and the process of training itself.

Risk assessment algorithms reveal facts like social inequalities affecting some ethnic or socioeconomic groups stronger than others.[250] The choice of the training procedure can have differential effects on different groups if the

242. If an algorithm is adjusted to minimize average errors it will fit the majority populations which will lead to different, and probably more, errors in a minority population. *See* Chouldechova & Roth, *supra* note 106, at 2; Irene Chen et al., *Why Is My Classifier Discriminatory?* (May 30, 2018), https://arxiv.org/abs/1805.12002 (arguing that issues of fairness should rather be addressed through data collection). For an instructive exemplification, *cf.* Karen Hao & Jonathan Stray, *Can You Make AI Fairer Than a Judge? Play Our Courtroom Algorithm Game*, MIT TECHNOLOGY REVIEW (Oct. 17, 2019), https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/ [https://perma.cc/E23R-DZYF].

243. Bornstein, *supra* note 226; Kleinberg et al., *supra* note 236, at 2 (discussing balance for the positive class, meaning the average score received by people possessing a property (likeliness for recidivism) should be the same in each group).

244. Bornstein, *supra* note 226; Kleinberg et al., *supra* note 243, at 2 (naming it the *balance for the negative class*).

245. *Cf.* Sam Corbett-Davies et al., *A Computer Program Used For Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*, THE WASHINGTON POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/3ZEB-MSUU].

246. Chouldechova & Roth, *supra* note 106, at 8; Laura Hudson, *Technology Is Biased Too. How Do We Fix It?*, FIVETHIRTYEIGHT (Jul. 20, 2017), https://fivethirtyeight.com/features/technology-is-biased-too-how-do-we-fix-it [https://perma.cc/35BF-HY56].

247. *See* Kleinberg et al., *supra* note 7, at 21-24; Kroll et al., *supra* note 195, at 680-82; Lehr & Ohm, *supra* note 22, at 703-04.

248. Kleinberg et al., *supra* note 7, at 22.

249. *Id.*; Kroll et al., *supra* note 195, at 680-81.

250. Kleinberg et al., *supra* note 7, at 26.

algorithm is not sufficiently optimized.[251] This can be avoided by rigorously testing the algorithm on new (testing) data and by cross-validation. Biases in algorithmic decisions can derive from the training data, especially if the training data includes human biases in previous (human) decisions.[252] A related problem arises if data exists only concerning certain groups; for example, black individuals are more likely to be arrested for a crime than white people, hence their conviction and imprisonment rate is higher, and they will appear to be more recidivistic.[253] The data fed into the algorithm can be constrained because not all past actions taken by the algorithm provide information on their outcome.[254] For example, recidivism can only be identified, if the defendant is released. To enable learning without this constraint (meaning learning about groups that typically are not granted parole), exploration can be undertaken by taking sub-optimal actions that allows us to collect new data.[255] This exploration in turn can lead to effects detrimental to individuals. Thus, it is ethically problematic. The risk of bias arises when past actions reinforce the choosing of the same action in the future in a feedback loop, for example, a place-based-policing algorithm may disproportionately concentrate police presence in areas with high arrest rates which in turn will keep arrest rates high.[256] To avoid over-policed communities, the policing intensity may be taken into account when recalibrating the algorithm, or incidents are only added to the training data if the arrested person is convicted.[257]

Algorithms can be designed to alleviate biases or, rather, the underlying statistical patterns.[258] For example, fairness in algorithms can be improved by "hiding" or removing sensitive information (*fair representation learning*).[259] Alternatively, *adversarial learning* can mitigate biases, for example by developing another ML algorithm that will be trained to minimize its ability to predict race.[260] Furthermore, the magnitude of discriminatory factors can be assessed through examining the ML algorithm, which might reveal the

251. *Id.* at 23-24.

252. Chouldechova & Roth, *supra* note 106, at 2, 6; Kleinberg et al., *supra* note 7, at 22 (recommending the use of an objective measure to predict the outcome).

253. Kleinberg et al., *supra* note 7, at 26.

254. Chouldechova & Roth, *supra* note 106, at 2; Kleinberg et al., *supra* note 7, at 20.

255. Chouldechova & Roth, *supra* note 106, at 2-3.

256. Chouldechova & Roth, *supra* note 106, at 6; Lepri et al., *supra* note 105, at 4; Selbst, *supra* note 61, at 135; Veale et al., *supra* note 64, at 7.

257. Chouldechova & Roth, *supra* note 106, at 7; Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACHINE LEARNING RES. 160, 166-67, 169 (2018).

258. Sam Corbett-Davies et al., *Even Imperfect Algorithms Can Improve the Criminal Justice System*, THE NEW YORK TIMES (Dec. 20, 2017), https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html; Chouldechova & Roth, *supra* note 106, at 7; Kroll et al., *supra* note 195, at 682-92; Lehr & Ohm, *supra* note 22, at 704-05; Lepri et al., *supra* note 105, at 616-18.

259. Chouldechova & Roth, *supra* note 106, at 7.

260. Alex Beutel et al., *Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations* (July 1, 2017), https://arxiv.org/abs/1707.00075 [https://perma.cc/V9AA-L5GK]; Harrison Edwards & Amos Storkey, *Censoring Representations with an Adversary* (Nov. 18, 2015), https://arxiv.org/abs/1511.05897 [https://perma.cc/M7DC-V9BL]; Brian Hu Zhang et al., *Mitigating Unwanted Biases with Adversarial Learning*, 2018 CONF. AI, ETHICS & SOC'Y, 335 (2018), https://arxiv.org/abs/1801.07593 [https://perma.cc/rxz5-cqcf].

direct use of illicit factors (e.g. race or sex). If the algorithm did not have access to such factors, simulated data can be used to experiment and compare the results of the ML algorithm with expected results.[261]

Prediction algorithms can be used to benefit disadvantaged groups ("disparate benefits"), for example by specifically detecting low-risk defendants with recidivism detection software.[262] To detect discriminating algorithms, of all the algorithms' components – including its training data – should be examined and audited.[263] Different approaches of evaluation and audit of ADM-software will be discussed in the next section.

C. ACCOUNTABILITY AND TRUST

Accountability of legal authorities means that they have to justify their actions or decisions and take responsibility in case of failures.[264] A first step towards accountability is transparency and enabling the public to evaluate the use of ADM-software. Through public scrutiny, the elected officials who decided to deploy ADM-software can be held responsible in case of failure.[265] However, public transparency may not suffice to control the use of ADM-software.

1. Enhancing Responsibility by Audit

For sensitive decision-making, especially when citizens and their fundamental rights are concerned, ADM-software should only be used if it allows for evaluation.[266] For this purpose, evidence is needed, both concerning the goals of the ADM-system, and its compliance with these goals.[267] Recording details of the decision-making process in an audit log facilitates assessing the validity of the decision.[268] Alternatively, the system's source code and the encoded policies can be disclosed to the public.[269] Others propose the disclosure of the ADM-software's assessment, for example, whether it sufficiently promotes transparency and accountability,[270] and whether it was thoroughly tested on a wide range of scenarios before its use,

261. Kleinberg et al., *supra* note 7, at 29.

262. *Id.* at 36 (applying to where the example of pre-trial release decisions based on a prediction of the failure to appear in court is presented). *See also* Anna Maria Barry-Jester et al., *Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?*, FIVETHIRTYEIGHT (Aug. 4, 2015), https://fivethirtyeight.com/features/prison-reform-risk-assessment/ [https://perma.cc/PB63-SQCZ].

263. Kleinberg et al., *supra* note 7, at 32, 39.

264. *See* Helen Nissenbaum, *Accountability in a Computerized Society*, 2 SCI. & ENGINEER. ETHICS 25, 26-27 (1996) (advocating for a strong culture of accountability and answerability in a computerized society).

265. Edwards & Veale, *supra* note 148, at 41.

266. Desai & Kroll, *supra* note 157, at 43.

267. *Id.*

268. *Id.* at 45.

269. Citron, *supra* note 5, at 1308-09. Risks are discussed, such as public safety concerns or a long process of reported flaws and corrections might be cost- and time-consuming. *Id.* Citron argues that public scrutiny concerning open-source software might be less expensive than reviewing closed software by third parties. *Id.*

270. *Id.* at 1308-10.

especially with regard to distorted policies and discriminatory biases.[271] A rather grassroots approach could be adopted by encouraging the general public and stakeholders to participate in the building of ADM-software.[272] To further accountability in the context of algorithmic decisions, Andrew Tutt proposed an administrative agency to monitor and regulate algorithms, parallel to the Food and Drug Administration (FDA) in the United States of America.[273] Such an agency could either set design standards and best practices, stipulate requirements of transparency, or it could act as a regulator giving pre-market approval.

The most pressing need is the determination of the requisite level of transparency.[274] Too much transparency could enable the malicious exploitation by strategically "gaming" the system.[275] Additionally, legitimate interests such as intellectual property rights relating to the ADM-software would be violated.[276] If training data is disclosed, infringements of privacy rights might be faced, for example if a dataset for training facial recognition software is examined with respect to diversity.[277] Yet, some form of verification of ADM-systems' functioning and security is indispensable[278] and will typically be done prior to the deployment in a non-public manner. The public interest and need for disclosure will not be satisfied through private or confidential software evaluation. At least key information such as metadata could be disclosed without jeopardizing the secrecy of the ADM-software itself.[279] The type and architecture of the ML algorithm (Random Forests or an ANN), can easily be reported. Often, open-source frameworks like Google's *TensorFlow* are used,[280] therefore revealing the type of algorithm and

271. Citron, *supra* note 5, at 1310; Edwards & Veale, *supra* note 148, at 76. Admittedly, this should be obvious. Yet Citron presents cases of automation software that was not adequately tested before its deployment. For testing of credit-scoring systems, *see* Citron & Pasquale, *supra* note 150, at 24-25.

272. Citron, *supra* note 5, at 1312 (proposing the establishment of information technology review boards, so that interested groups can comment on the system's design and testing).

273. *Cf.* Tutt, *supra* note 162, at 119-22; Michael Segal, *We Need an FDA For Algorithms*, NAUTILUS (Nov. 1, 2018), http://nautil.us/issue/66/clockwork/we-need-an-fda-for-algorithms [https://perma.cc/RDF2-DMQS].

274. *See* Kartik Hosanagar & Vivian Jair, *We Need Transparency in Algorithms, But Too Much Can Backfire*, HARVARD BUSINESS REVIEW (Jul. 23, 2018), https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire [https://perma.cc/PNK6-5FG8].

275. Kroll et al., *supra* note 195, at 658.

276. Burrell, *supra* note 158, at 3-4; *see also* Desai & Kroll, *supra* note 157, at 40; Edwards & Veale, *supra* note 148, at 65; Zalnieriute et al., *supra* note 56, at 14.

277. Kroll et al., *supra* note 195, at 658.

278. *Id.* at 662-65 (outlining software verification (through which one proves mathematically that software has certain properties), and other aspects of evaluation like cryptographic commitments).

279. Some administrative bodies already published information on the ADM-software they used or invited journalists and other interested parties to explain their ADM-tools. *See* Veale et al., *supra* note 64, at 6.

280. *See, e.g.*, Alison DeNisco Rayome, *The 10 Most Popular Machine Learning Frameworks Used by Data Scientists*, TECHREPUBLIC (Sep. 14, 2018), https://www.techrepublic.com/article/the-10-most-popular-machine-learning-frameworks-used-by-data-scientists/ [https://perma.cc/8YUF-24DR]; Sara Bertram, *Ene, Mene, Muh – Und Raus Bist Du*, iX HEISE MAGAZINE 66 (2019), https://www.heis e.de/select/ix/2019/1/1545999823788057 [https://perma.cc/SES3-UJM8] (Ger.). *Cf.* Kartik Hosanagar, *The Democratization of Machine Learning: What It Means for Tech Innovation*, KNOWLEDGE@WHARTON (Apr. 13, 2017), https://knowledge.wharton.upenn.edu/article/democratization-ai-means-tech-innovation/ [https://perma.cc/6Q7U-KPNW].

the architecture would not infringe on any secrecy concerns. Furthermore, instead of disclosing the whole training data, one could publish metadata of interest. For example, with facial or dialect recognition, the composition of the different dialects, genders, age groups, and ethnicities can be released without violating privacy rights of data subjects. The process of collating metadata on training data could already elicit self-criticism and improvements of datasets.

2. Issues of Legitimacy and Separation of Powers

According to the principle of separation of powers, legislative, executive, and judicial power should be kept separate, meaning three distinct entities pass laws, enforce them, and apply them to cases if disputes arise.[281] This clear division blurs with the use of ADM. The algorithm's designer, the authority putting a query to the system, and the system outputting a decision, all represent an equivalent of some domain of state authority.[282] The substance of rules might change when they are translated into code.[283] Its rules will be applied in a distorted manner, law will effectively be changed. These new rules, or their flawed application, are not overt and might not be detected by the relevant administrative authority, meaning they cannot review the code, and the new legal practice cannot be debated by the public.[284] The divergent implementation of rules will mostly be caused by accident, not by wielding politically legitimized power.[285] Public officials have a certain authority to determine the application of rules and are legitimized through elections or appointment. They are accountable to the people, but are hardly proficient enough to review ADM-software and bestow legitimacy to them and their applying law.[286] The programmers on the other hand do not have the authority and legitimacy to establish new rules or standards of applying rules already in force.[287] Because the software developers work anonymously and without mandate, citizens may perceive them as shadowy and not trustworthy. The legitimacy of algorithmic decisions will be weakened. This problem may arise also in other cases of public commission, but the opacity of ML models complicates the scrutineering of third-party software. Preemptive rulemaking procedures that precede the automatization of decision-making can restrict law-making through programming and safeguard the separation of powers.[288] For example, government institutions and legislators may issue interpretative rules or policy statements which can be publicly discussed.[289]

281. *Separation of powers*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/topic/separation-of-powers [https://perma.cc/M4FJ-VQ5T] (last visited Mar. 27, 2020).

282. Crawford & Schultz, *supra* note 150, at 120.

283. Citron, *supra* note 5, at 1254-55.

284. *Id.*

285. *Id.* at 1297.

286. *Id.* at 1295.

287. *Id.* at 1254-55.

288. *Id.* at 1312-13.

289. *Id.*

3. Trusting the "Human Element"

If ADM-software is deployed, issues of human-machine-interaction arise. The judge or police officer using decision-support software might trust its recommendation or result and fall victim to the *Automation Bias*.[290] Even if a person contests a decision in a hearing, if he presents all his arguments, or if a review of an algorithmic decision takes place, the legal authority might not be truly impartial. Because judges reviewing algorithmic decision cannot grasp the reasoning of the system, a meaningful review of the case through a *human in the loop* will most likely not take place.[291] Instead, decision-makers might rely too much on the result of the ADM-software and justify the results *ex post*.[292] This failing might be remedied by educating decision-makers about the biases and fallacies automated decision-making entails.[293] Alternatively, the ADM-software might provide a prioritized list of options, so that the decision-maker can exercise discretion based on experience and intuition.[294] Decision-support software can assist legal decision-makers with tasks easily automated, so they can focus on more advanced work.[295]

4. Trustworthy Algorithms?

Building trust in ADM-software seems questionable at first. However, if the software is carefully designed and reviewed before its deployment, and if collective awareness and general knowledge concerning automated decision-making is raised, people might be able to trust an algorithm more than a human. Because algorithms obtain decisions by rule-based processes, they do not have any motives. Hence, there would be no grounds for distrust.[296] Furthermore, they might improve legal certainty through automation.[297] Yet, ensuring legal certainty can impede legal evolution based on changing social conventions or values.[298] ADM-tools should be reviewed on a regular basis and adjusted to changes.[299] A correctly adjusted decision-making system plainly applying legal rules or identifying relevant facts of the case, would certainly be more trustworthy than a human "black box."

290. Citron, *supra* note 5, at 1254; Oswald, *supra* note 146, at 16-17; Veale et al., *supra* note 64, at 4, 9.

291. Citron, *supra* note 5, at 1298.

292. Even if the decision-maker does not accept the ADM-software's outcome or suggestion, the *anchoring effect* might influence the decision-maker in her judgment. *See* Kahneman, *supra* note 170, at 119-28. The initial suggestion rendered by the ADM will cause her unconsciously to align her decision with this result.

293. Edwards & Veale, *supra* note 148, at 76.

294. Veale et al., *supra* note 64, at 4, 9.

295. *Id.* at 8.

296. Brian Sheppard, *Warming Up to Inscrutability: How Technology Could Challenge Our Concept of Law*, 68 U. TORONTO L. J. 36, 53-54 (2018) (discussing the necessity of public officials having a critical reflective attitude).

297. Oswald, *supra* note 146, at 6.

298. Sheppard, *supra* note 296, at 54-55 (arguing that law would become "static").

299. *See* Veale et al., *supra* note 64, at 7-8 (discussing examples of ADM-software that are required to be adjusted to changing circumstances).

D. RIGHT TO BE HEARD AND RIGHT TO NOTICE

If ADM-software is deployed, the importance of hearings increases.[300] The affected individual will perceive the algorithmic decision as fairer, if he had an opportunity to be heard. The research on procedural justice seems to suggest that people care about being heard and having their arguments considered. Presenting their argument and winning because of it is of less importance. When legal authorities explain their decisions, they should acknowledge the concerns of the affected parties and show that they were taken into account, even if an adverse decision was reached.[301] Giving people the opportunity to voice their view and arguments before an ADM-system is challenging. As of now, a system analyzing parties' statements and weighing their arguments based on semantic understanding of their writings has yet to be developed. People will not really feel heard, if they are facing a form with boxes to check. The judges' role entails interactive aspects such as communicating to the parties, moderating the legal dispute, explaining the applicable law and the reasons of their decisions, and contributing to the education of civil society.[302] In this respect, a human in the loop would be needed to listen to the arguments made and to consider whether they have relevance to the decision. The judge holding a hearing should be trained concerning ADM, their fallibilities, and their own proneness to automation bias.[303] Furthermore, if the judge relies on the algorithmic decision, he should explain so in detail and provide the reasons for the algorithmic decision.[304] Lastly, if citizens have the possibility to clarify facts and expound their views, erroneous decisions can easily be detected.

Moreover, citizens should receive notice if a decision was reached and ADM-software contributed to it. [305] Automated decision-making can curtail due process if the system adjudicates in secret, so the affected person does not know about the decisions rendered, or if the records of automated decisions are not kept, so that a review of the considerations leading to the decision cannot be reviewed.[306] Audit trails capturing the facts and rules that influenced the decision, and the administrative authorities may provide individuals with notice of the automated decision.[307] This would allow for a more apt judicial review. The automation bias can be countered, as judges

300. Edwards & Veale, *supra* note 148, at 76; Crawford & Schultz, *supra* note 150, at 126-27 (in the context of private actors, proposing a data arbiter as a third party to settle concern about algorithmic decisions).

301. Tyler, *Procedural Justice and the Courts*, *supra* note 121, at 31.

302. Sourdin, *supra* note 55, at 1118, 1124 (mentioning that the interactive nature of the judge's role would change if "judicial AI" was used to adjudicate legal disputes).

303. Citron, *supra* note 5, at 1306.

304. *Id.* at 1307.

305. Edwards & Veale, *supra* note 148, at 76; Crawford & Schultz, *supra* note 150, at 125; Citron & Pasquale, *supra* note 150, at 28 (discussing in the context of private companies and proposing a right to petition providers).

306. Citron, *supra* note 5, at 1253.

307. *Id.* at 1305. With regard to private providers, *cf.* Crawford & Schultz, *supra* note 150, at 127-28.

might assess algorithmic decisions more critically if they are exposed to their finding.[308]

E. DIGNITY AND RESPECT

Citizens want to be treated respectfully by the authorities. One component of a respectful treatment is politeness and sincerity – characteristics that only humans can fulfil for now.[309] Conversely, an ADM would not interact with the concerned people directly and if so, it would at least not be impolite.[310] Respectful treatment might be less problematic in automated decision-making procedures. However, since most ADM is concerned with decision-support or providing instructive information for the actual (human) decision-maker, the essential factor of respectful treatment is the human in the loop: the judge deciding whether to grant asylum, or the police officer patrolling a crime hotspot.

Closely related is the question whether the individual's dignity is respected. Algorithmic decision about humans might be considered undignified. The human would be objectified and dehumanized in algorithmic processes.[311] As a consequence, a human-in-the-loop approach can be demanded, so only a human will make a final decision.[312] Thus, assigning responsibility and accountability for a decision is less problematic.[313] Nevertheless, the efficacy of human review has to be questioned.[314]

Additionally, dignity can be safeguarded by rigorously enforcing individual rights, especially fundamental rights and data protection law.[315] Mass surveillance can be conducted more effectively and more efficiently with ML since it enables the automated analysis of huge amounts of data – be it text, sounds, or images. Mass surveillance systems can be used to monitor whole populations, leading to chilling effects and deterring citizens

308. Citron, *supra* note 5, at 1305-06.

309. For the latest progresses in Emotion AI, *see* Charles Towers-Clark, *Making AI More Emotional – Part One*, FORBES (Jan. 28, 2019), https://www.forbes.com/sites/charlestowersclark/2019/01/ 28/making-ai-more-emotional-part-one/#6562b9255fc1 [https://perma.cc/KCY2-UMZ7]; Charles Towers-Clark, *Making AI More Emotional – Part Two*, FORBES (Jan. 31, 2019), https://www.forbes.com /sites/charlestowersclark/2019/01/31/making-ai-more-emotional-part-two/#2b943fc0318e [https://perma.cc/C8BA-CL2J].

310. If the ADM comprised a personal-assistant-like component, it would appear as well-mannered as C-3PO. Otherwise, one could argue that a simple, functional result is asocial in the true sense of the word.

311. ADM could be perceived as the *datafication* of individual. The term "datafication" was coined in VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA 73-97 (2014).

312. Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216, 230-32 (2017); *see also* GDPR, art. 22.

313. Moreover, the authorship of a decision would be unequivocal. *See* Sourdin, *supra* note 55, at 26-27.

314. *See supra* discussion in section *D. Right to be Heard and Right to Notice.*

315. *See* Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321, 1374-79 (1992). Fundamental rights and data protection law cannot be addressed here. Although the adherence to individuals' rights benefits the perception of fairness and legitimacy of legal authorities, it is an issue of legislation and regulation rather than a matter of procedure. It will be presumed that authorities act lawfully.

from exercising their rights.[316] Living in such a panopticon certainly impairs human dignity. Safeguards against mass surveillance already exist in many legal systems but may lack appropriate restrictions and oversight.[317] The respective laws should be adjusted and enhanced to address AI tools, especially the looming ubiquity of facial recognition in public spaces.

In contrast to perceiving ADM as a threat, algorithmic decisions can be regarded instrumentally by focusing on the question whether algorithmic decision-making will outdo a human in achieving a task or a goal.[318] Algorithmic decisions on humans might not be considered undignified, but crucial to obtain the optimal, the most neutral as well as the time- and cost-efficient[319] result. If substantive justice can be improved by deploying ADM-software, and if legal proceedings will be less long-winded and tedious, the citizens affected might be better off. So long as algorithms do not determine our values and goals, the deployment of ADM-software might be decided based on competency.[320] And "if the delegation of decision making power is carried out responsibly, we may be creating a much more humane society. Some of the most humanistic decisions may well come from decision makers which are not human."[321] In this case, the fundamental question will be about how to assign responsibility for algorithmic decisions. ADM can only be causally[322] responsible, but not legally, politically, or morally.[323] In the long run, humankind must decide in which cases and to what extent algorithms may render decisions affecting humans.

CONCLUSION

Legal decision-making is already being automated, and this trend will expand to more areas of application. Media and academia are torn between marveling at the increasing capabilities of ML algorithms and succumbing to the anxiety of their opacity and unmanageable complexity. Citizens have even fewer opportunities to understand these intricate algorithms and how their results were obtained. If they are faced with legal authorities deploying ADM-software, citizens might succumb to the feeling of powerlessness, with detrimental effects to their perception of legitimacy and to their willingness to comply with decisions. Even if legal authorities take suggestions of decision-support systems into consideration in their final decision, the

316. *See, e.g.*, Jonathon W. Penney, *Chilling Effects: Online Surveillance and Wikipedia Use*, 31 BERKELEY TECH. L.J. 117, at 147-53 (2016).

317. *E.g.*, the European Court of Human Rights found that "the intelligence services of the United Kingdom take their Convention obligations seriously and are not abusing their powers" but identified a "lack of oversight of the entire selection process" and "the absence of any real safeguards applicable to the selection of related communications data for examination." Big Brother Watch and Others v. the United Kingdom, no. 58170/13, §§ 387-88, ECHR 2018. *See also id.* §§ 314-20, 328-47, 355-57.

318. James Moor, *Are There Decisions Computers Should Never Make*, 1 NAT. & SYS. 217, 227 (1985).

319. The advantage of cost-efficiency is obvious. Time-efficiency, on the other hand, is often underappreciated. Especially when court proceedings last several months or even years.

320. Moor, *supra* note 318, at 225-28.

321. *Id.* at 228. As already pointed out, algorithmic decision-making could improve equality.

322. The term "causal" is used in the scientific context.

323. *Id.* at 227-28.

impartiality and open-mindedness of the decision-maker can and will be questioned. Citizens might perceive the authority's decision as even more opaque and imponderable than the decisions of a human, despite the inscrutability of human decision-making. They will experience algorithmic decision-making to be Kafkaesque.

Hence, trust in automated decision-making needs to be enhanced. This sublime goal can only be achieved, if algorithmic decisions are transparent and fair. The reasons and influencing factors of algorithmic decisions need to be disclosed to citizens, as is necessary with human decisions. The quality of ADM-software must be ascertained and ascertainable, especially with respect to possible biases and discrimination. Affected citizens should be empowered to take action and challenge algorithmic decisions that are deemed to be flawed, incorrect, or unlawful. Citizens need to be notified that decisions concerning them were made by using ADM. In review proceedings, the authorities must consider the arguments presented by the parties. The decision-maker or reviewer should be wary of algorithmic decisions or recommendations, and only adopt them if he independently ensured their validity. Lastly, the affected citizens should be treated respectfully and with dignity. Since ADM cannot assume such social duties, legal authorities should pay special attention to the communicative aspects of their roles. In the future, the focus of legal decision-making might shift from the solitary analysis done by the judge to hearings allowing parties to present their arguments. Ultimately, legal authorities might inform people about the ADM-software that was used and explain their functioning and rationale to build trust in and improve the acceptance of algorithmic decisions. It does not suffice to safeguard individuals' interests in ADM-processes. The general public as sovereign needs to be able to ascribe accountability for algorithmic decision-making. The development and deployment of ADM-software must not lead to public officials shirking away from responsibility. The development of ADM, especially the "translation of code into law" should be guided by democratically legitimized directives. To promote public scrutiny of automated decision-making, key information on deployed ADM-software, such as metadata on the ML model and the training data, must be made available for public debate. Using open-source software to build the ADM-model can enhance trust and auditability.

The decisive role in automated decision-making (at least in the near future) will be played by the humans-in-the-loop who review, control, and communicate algorithmic decisions. As long as the humans-in-the-loop ensure the respectful treatment and the consideration of the arguments of people concerned by the decision, the use of ADMs can actually improve the perception of fairness and procedural justice. The human element (still, or now more than ever) serves important functions. The focus of judges might shift from rigorously applying the law to explaining the decisions and its reasons to the parties. Judges could spend more time on listening to arguments and the "story" of the people appearing before court. Police officers could spend more time communicating to people to build trust and improve

compliance. Certainly, ADM may threaten procedural justice. However, they also open up possibilities to improve the accuracy, efficiency, and neutrality of legal decision-making, which may foster trust in legal authorities and compliance with their decisions.

The remaining question is, whether AI deciding on humans encroaches on human dignity. Ultimately, this issue is for humanity to decide. Since we have already begun to use AI for legal decision-making, we (unwittingly) started the process.