Legal Priorities Project

The challenges of artificial judicial decision-making for liberal democracy

Christoph Winter

LPP WORKING PAPER N° 3-2021

In: P. Bystranowski, P. Janik, & M. Próchnicki (Eds.), *Judicial Decision-Making. Economic Analysis of Law in European Legal Scholarship,* vol 14. Springer, Cham. (2022).
https://doi.org/10.1007/978-3-031-11744-2_9

The Challenges of Artificial Judicial Decision-Making for Liberal Democracy

*Christoph Winter**

March 17, 2021

ABSTRACT

The application of artificial intelligence (AI) to judicial decision-making has already begun in many jurisdictions around the world. While AI seems to promise greater fairness, access to justice, and legal certainty, issues of discrimination and transparency have emerged and put liberal democratic principles under pressure, most notably in the context of bail decisions. Despite this, there has been no systematic analysis of the risks to liberal democratic values from implementing AI into judicial decision-making. This article sets out to fill this void by identifying and engaging with challenges arising from artificial judicial decision-making, focusing on three pillars of liberal democracy, namely equal treatment of citizens, transparency, and judicial independence. Methodologically, the work takes a comparative perspective between human and artificial decision-making, using the former as a normative benchmark to evaluate the latter.

The chapter first argues that AI that would improve on equal treatment of citizens has already been developed, but not yet adopted. Second, while the lack of transparency in AI decision-making poses severe risks which ought to be addressed,

* Assistant Professor, Instituto Tecnológico Autónomo de México (ITAM), Faculty of Law
Visiting Scholar, Harvard University, Department of Psychology
Director, Legal Priorities Project
Email: christoph_winter@fas.harvard.edu

I am grateful to Suzanne Van Arsdale, Julian Aveling, Cullen O'Keefe, Antonia Juelich, Nick Hollman, Jonas Schuett, Leonie Koessler, Renan Araújo, Eric Martínez and the participants at related presentations at Harvard Law School and the IVR World Congress 2019 in Lucerne for their comments and inspirations on this Article. I would also like to thank the editors of this book for their very constructive feedback.

AI can also increase the transparency of options and trade-offs that policy makers face when considering the consequences of artificial judicial decision-making. Such *transparency of options* offers tremendous benefits from a democratic perspective. Third, the overall shift of power from human intuition to advanced AI may threaten judicial independence, and with it the separation of powers. While improvements regarding discrimination and transparency are available or on the horizon, it remains unclear how judicial independence can be protected, especially with the potential development of advanced artificial judicial intelligence (AAJI). Working out the political and legal infrastructure to reap the fruits of artificial judicial intelligence in a safe and stable manner should become a priority of future research in this area.

Keywords: Judicial decision-making, artificial intelligence, liberal democracy, discrimination, transparency, judicial independence, separation of powers, advanced artificial judicial intelligence (AAJI)

CONTENTS

The application of artificial intelligence (AI) to judicial decision-making has already started. Significant progress has been made, not only in the United States, most prominently with regard to bail decisions,[1] but also in Russia[2] and Mexico.[3] China has placed more than 100 robots in courts offering legal advice to the public,[4] and Estonia is piloting a program in which small scale civil suits are decided by an algorithm.[5] Furthermore, a recent survey suggests that legal scholars believe that, on average, almost 30% of judicial decision-making will be carried out by AI in only 25 years' time, tripling their estimate that the current role of AI accounts for less than 10% of judicial decision-making.[6] Against this background, it seems plausible to assume that future advances in AI will revolutionize the judicial sector. To many, AI not only promises greater fairness, justice, and legal certainty, but it may finally satisfy the legal requirements of the internationally accepted concept of a *fair trial*, in particular with regards to access to justice, as recognized by, among others, Article 6 of the European Convention on Human Rights (ECHR), Article 10 of the Universal Declaration of Human Rights (UDHR), and Articles 7 and 25 of the African Charter on Human and Peoples' Rights (ACHPR).[7]

At the same time, issues of discrimination and transparency have emerged in the context of bail decisions, putting liberal democratic principles under pressure. Although intense debates about the criminal justice system and demands for what has been coined "algorithmic fairness" have been ushered, further challenges that arise from the shift in authority from human intuitions to artificial intelligence within the judicial system remain unclear and neglected in the discourse. As of now, there has been no systematic analysis of the risks to liberal democracy from implementing AI into judicial decision-making. This article therefore sets out to fill this void by identifying and engaging with challenges arising from artificial judicial decision-making.

---

1 Angwin et al. (2016).

2 Zavyalova V (2018).

3 In Mexico, the Expertius system is advising judges and clerks regarding the determination of whether the plaintiff is or is not eligible for pension. *See* Carneiro D et al (2015).

4 World Government Summit (2018). The robot named Xiao Fa has a vaguely humanoid appearance and provides simple legal advice, such as how to bring a lawsuit or retrieve case histories, verdicts, and laws.

5 Niiler, E (2019).

6 Martinez and Winter (2021). As of January 8th, 2021, 307 legal scholars based in the United Kingdom, India, New Zealand, Bangladesh, Australia, Canada, and South Africa have responded to the relevant questions of the survey.

7 For various rights associated with a fair trial, *see also* the Sixth Amendment to the United States Constitution.

In this way, this analysis not only offers a first attempt to systematize these challenges, but it also outlines a number of crucial issues which demand further research. Throughout the analysis, I will apply a comparative perspective between human thought processes and artificial intelligence,[8] with an emphasis on ethical issues arising from (I.) large-scale discrimination and (II.) a potential lack of transparency within the judiciary. Indeed, I will argue that these issues pose some short- and medium-term threats. However, drawing most notably on recent research by Kleinberg and colleagues concerning bail decisions,[9] I am confident that related technical and ethical issues can be solved to yield outcomes better than those from human decision-making, from a wide range of philosophical perspectives on discrimination. While the lack of transparency in AI decision-making, narrowly construed, does pose risks which ought to be addressed, AI can also increase transparency in another – arguably more important – domain, which I will refer to as *transparency of options*. The far greater long-term threat for liberal democratic values lies in (III.) the overall shift of power from human intuition to advanced AI and, more precisely, in the possibly accompanying threats to judicial independence and the separation of powers. Accordingly, an AI-based judiciary might significantly contribute to the rise of digital authoritarianism.

Before engaging with these questions, it is necessary to clarify and limit the scope of this analysis. A common framework within the discourse is the distinction between whether AI will *supplement* or *replace* human beings.[10] Will AI help human judges to detect biases or make better decisions, or will it replace human judges altogether? Furthermore, the discussion in law and policy is usually focused on *narrow* applications of so-called AI that use machine learning or simple decision trees to supplement or partially replace human decision-making. I will concentrate on these partial replacements and corresponding issues of discrimination and transparency in the first two sections of this chapter. However, in the final section I will move on to more advanced forms of AI that could replace the vast majority of human judicial decision-making, thereby going beyond present applications, possibilities, and the main focus of literature. Even though this contemplates advanced forms of AI understood as significantly exceeding current capabilities, one should note that this would not require superintelligence that "greatly exceeds the cognitive performance of humans in virtually all domains of interest."[11] Instead, as

---

8 *Note* that this approach, as intuitive as it may sound, is not that common. Indeed, most critics of the application of AI towards the judicial sector focus on the disadvantages of present AI while overlooking the tremendous shortcomings of human decision-making.

9 Kleinberg et al. (2018); Kleinberg et al. (2019).

10 *See,* among others, Sourdin (2018).

11 *See generally*, Bostrom (2014); in a similar vein, Bostrom (2006), p.11.

I will explain later, it would require the development of an *advanced artificial judicial intelligence* (AAJI) with more limited domain knowledge and capabilities.

A prevailing argument among researchers and politicians alike seems to indicate that one should focus on present issues rather than trying to solve those which *may* occur in the future.[12] However, research about future developments, such as the implementation of more advanced forms of AI into the judiciary and its entailing risks is crucial because outlining future risks may provide valuable information on what sort of developments ought to be followed with particular care. If, for instance, great risks for liberal democracy are likely to occur in the future as a result of the application of AI, one may try to find ways in the present to mitigate such risks, if one considers liberal democracy to be a desirable political system. To put it simply: it can be beneficial to avoid an encounter with a predator in the first place, rather than trying to run away when it is already in front of you.

I. FROM HUMAN BIAS TO ALGORITHMIC FAIRNESS

This analysis will first focus on equal treatment for all citizens which, arguably, is *the* distinctive part of *liberal* democracies.[13] Much of the academic[14] and popular[15] debate on AI in the judiciary has been dominated by issues of illegitimate discrimination based on sex or race. This is understandable, given that Google's face recognition algorithm labeled black people as gorillas,[16] algorithms employed in job selection decisions favored white males,[17] and white people benefitted more than

---

12 *See*, e.g., Reiling (2018): "Don't spend time thinking about things that don't exist yet."; *Cf* also Sourdin and Cornes (2018), p. 113, who seem to classify such considerations as "unhelpful". While I agree that, for the foreseeable future, "human intelligence may (rather) be supplemented by technological advances", it seems almost naïve to consider only the more plausible scenarios and reject consideration of how one should act in case less likely scenarios unfold. Instead, such concerns and related research seem well justified not only from an expected value perspective, but from any reasonable version of the precautionary principle. *Cf.* also the emerging argument on bridging near- and long-term concerns about AI: Baum, 2018, 2020; Cave and Ó hÉigeartaigh, 2019; Prunkl and Whittlestone, 2020.

13 Mukand and Rodrik (2020).

14 *See*, among others, Chander (2017); Chen (2019); Hacker (2018); Kleinberg et al. (2019); Sourdin (2018), *supra note* 5, at 1128-1129.

15 *See*, *e.g.*, Levin (2016); Smith (2016).

16 Zhang (2015).

17 *See*, *e.g.,* the well-documented case regarding application for a medical school in the UK: Lowry and Macpherson (1988).

black people from the COMPAS[18] recidivism algorithm,[19] among others. It is clear that improvements are needed. Yet, it is a separate question whether these issues justify the abolishment or omission of incorporation of AI into the judicial sector. After all, the much longer history of human intelligence (HI) in the judiciary seems to indicate that human decision-making has not done justice to all groups either. While it may be possible to mitigate discrimination resulting from implicit human biases to some degree,[20] we need to identify how feasible it is for AI to overcome such biases compared to HI. If there is a sufficiently good chance that AI is even better positioned to do so in the long term, one should not reject entirely the idea of artificial judicial decision-making due to discrimination, but rather concentrate on how to mitigate discrimination by AI. In order for AI to play an important role in the judiciary of the future, it does not have to be perfect, but only *better* than HI. Indeed, we may even be obligated to adopt it in such cases.

Pertaining to discrimination, the law works remarkably similar across jurisdictions in distinguishing between *direct* and *indirect* discrimination, also understood as disparate *treatment* and disparate *impact* respectively.[21] Direct discrimination is typically understood as treating a person less favorably than another person in a comparable situation on the basis of any protected ground, such as sex, racial or ethnic origin, religion, disability, age, or sexual orientation.[22] Hence, the disadvantageous treatment is based on the possession of specific characteristics. Here, cases turn on the analysis of the causal link between the protected ground and the less favorable treatment,[23] as well as the comparability of persons in similar

---

18 COMPAS stands for "Correctional Offender Management Profiling for Alternative Sanctions" and was developed by Northpointe (now Equivant). The COMPAS recidivism algorithm is used by U.S. courts in many states to assess the likelihood of a defendant becoming a recidivist.

19 *See* Larson et al. (2016).

20 *See*, among others, Carnes et al. (2015); Lai et al. (2014); Devine et al. (2012).

21 *See*, for instance, Article 2 (2) of the EU Racial Equality Directive; ECtHR, Biao v. Denmark [GC], No. 38590/10, 24 May 2016, para. 89-90. For the US approach, *see* definitions provided in McGinley (2011), p. 626.

22 *See*, e.g., the very similar definitions offered under EU Law by Article 2 (2) of the EU Racial Equality Directive, by the case law of the ECtHR in ECtHR, Biao v. Denmark [GC], No. 38590/10, 24 May 2016, para. 89; and ECtHR, Carson and Others v. the United Kingdom [GC], No. 42184/05, 16 March 2010; and the US approach in Bolling v. Sharpe, 347 U.S. 497, 499, 74 S. Ct. 693, 694, 98 L. Ed. 884 (1954); Brown v. Board of Education, 347 U.S. 483 (1954); Washington v. Davis, 426 U.S. 229, 239, 96 S. Ct. 2040, 2047, 48 L. Ed. 2d 597 (1976). Some jurisdictions, e.g. German Basic Law Article 3, favor a non-exhaustive list of these grounds which means that further grounds can be added if deemed necessary.

23 To establish such a causal link, one must answer the following question: "Would the person have been treated less favorably had they been of a different sex, race, age, or in

situations. In some jurisdictions, direct discrimination can be justified,[24] in others it cannot.[25]

Indirect discrimination takes place when an apparently neutral criterion, provision, or practice would put persons protected by some general prohibition of discrimination (e.g., those having a protected characteristic) at a disadvantage compared to others.[26] If such discrimination occurs, it will have to be justified, i.e. the provision in question will be upheld only if it has a legitimate aim, and the means of achieving that aim is necessary and appropriate. In contrast to direct discrimination, indirect discrimination is based on apparently neutral criteria which are not formally prohibited. Nevertheless, the consequence of both direct and indirect discrimination is essentially the same: an individual who belongs to a protected group is disadvantaged.

Despite the fact that the application of these definitions can be tricky depending on the exact facts of the case and the evidence available, they work rather well in the context of algorithms, as they explicitly state what criteria are directly taken into account and which ones are not. For example, in the context of bail decisions, algorithms which do not directly take race, sex, religion, or other prohibited factors into account for assessing the flight risk of an individual do not discriminate directly.[27] However, as mentioned earlier, the COMPAS recidivism algorithm had worse effects for people of color in comparison to white people. Thus, the algorithm may indirectly discriminate by taking into account apparently neutral criteria such as the zip code, level of education, or a poor credit rating when predicting the flight

---

any converse position under any one of the other protected grounds?" If the answer is "yes", then the less favorable treatment is caused by the grounds in question.

[24] *See*, for instance, for the United Kingdom, Professor John Pitcher v. Chancellor, Masters and Scholars of the University of Oxford and Saint John the Baptist College in the University of Oxford [2019] UK Employment Tribunals 3323858/2016. Even for the most protected categories in the US like race and religion, direct discrimination is justifiable if it passes strict scrutiny. Note, however, that this is rarely the case, Fisher v. Univ. of Texas at Austin, 570 U.S. 297, 310, 133 S. Ct. 2411, 2419, 186 L. Ed. 2d 474 (2013).

[25] E.g., in the UK, direct discrimination can only be justified regarding age and disability. *See* United Kingdom, Equality Act 2010, Part 2 Chapter 2 Sections 13 and 19.

[26] United Kingdom, Equality Act 2010. *Cf.* also the case law of the ECtHR: ECtHR, D.H. and Others v. the Czech Republic [GC] (No. 57325/00), 13 November 2007, para. 184; ECtHR, Opuz v. Turkey (No. 33401/02), 9 June 2009, para. 183; ECtHR, Zarb Adami v. Malta (No. 17209/02), 20 June 2006, para. 80.

[27] If not the sole determent for accessing whether pretrial release is to be allowed, the flight risk is in many jurisdictions a crucial factor. *See*, for instance, State of New Hampshire v. Christina A. Hill (2019) Supreme State Court NH 2018-0637; in other states in the US, the likelihood that the defendant will be arrested for or convicted of a crime also matters, *cf.* Dabney et al. (2017), p. 408; Karnow (2008), p. 1.

risk. Since the abovementioned factors have often been influenced by previous discrimination,[28] the common worry that AI may *perpetuate* discrimination, even if the algorithm does not explicitly take race or sex into account, is justified.

However, the COMPAS case, which has strongly shaped the public discourse on AI in the judiciary, has limited value for informing long-term policy, as the exact algorithm is undisclosed and kept as a trade secret.[29] While the case shows that it is certainly possible to build algorithms that indirectly discriminate against protected groups (and perhaps discrimination is even likely, absent safeguards in development and application), our primary concern is whether it is possible to build an algorithm or AI that is beneficial from different philosophical points of view in comparison to human decision-making, even with respect to issues of discrimination. In this vein, recent research by Kleinberg and colleagues shows that such improvements are not merely a utopian vision, but feasible to build today.[30] Again focusing on bail decisions, the machine learning algorithm developed by Kleinberg and colleagues, which used gradient-boosted decision trees, was trained with a large dataset of 758,027 defendants who were arrested in New York City between 2008 and 2013. The dataset included the defendant's prior rap sheet, the current offense, and other factors available to judges for making the decision. When tested on over 100,000 different cases, the algorithm proved to be significantly better than human judges at predicting whether defendants would fail to appear or be re-arrested after release, which – depending on one's policy preferences – can come with diverging yet strong benefits. More precisely, simulations showed failure to appear and re-arrest ("crime") reductions of up to 24.7% and no less than 14.4% with no change in jailing rates, or jailing rate reductions up to 41.9% and no less than 18.5% without any increase in crime rates.[31] Crucially from the perspective of preventing discrimination, the algorithm was able to achieve crime reductions while simultaneously *reducing* racial disparities in all crime categories.[32] The adoption of the algorithm would thus put policymakers in the rather comfortable position of choosing between the options of releasing thousands of people pre-trial without adding to the crime rate *or* preventing thousands of crimes without jailing even one additional person – whilst reducing racial inequalities. Needless to say, these are not the only options, but they illustrate what kind of trade-offs policymakers will have to make with regards to the

---

28 *See* discussions of structural racism, among others, Wallace et al. (2017), Hammer (2018).

29 *Cf.* Chohlas-Wood A (2020). I will discuss the apparent lack of transparency in more detail in Section II.

30 Kleinberg et al (2018).

31 Kleinberg et al. (2018), p. 241.

32 Kleinberg et al. (2018), p. 237-238.

balancing of crime and detention rates, as well as the disproportionate imprisonment of minorities, particularly black and Hispanic males in the United States, if AI replaces HI. In fact, Sunstein argues that the algorithm developed by Kleinberg and colleagues "does much better than real-world judges (…) *along every dimension that matters*".[33]

Importantly, these results are not unique to New York City, as Kleinberg and colleagues were also able to obtain qualitatively similar findings in a national dataset.[34] However, one might be tempted to argue that it remains uncertain whether these results can be achieved in jurisdictions outside the US. Could it be that there is just something off with judicial decision-making in the US? Might there be a bug in the system which does not exist in other jurisdictions when it comes to bail decisions? After all, common and civil law systems vary,[35] and legal education in the United States is fundamentally different from most places, even other common law systems such as Australia, the United Kingdom, or India. Despite this, I fail to see why such differences would lead to significantly different outcomes because, first of all, the law works remarkably similar with regards to bail decisions.[36] Second, Sunstein points out that the analysis of the data suggests that cognitive biases can explain why AI so clearly outperforms HI.[37] This is a relevant and important insight given that biases occur worldwide irrespective of the jurisdiction.[38] To be precise, Sunstein argues that judges make two fundamental mistakes which significantly influence their overall performance.[39] They treat high-risk defendants as if they were low-risk when their current charge is relatively minor, and they treat low-risk people as if they were high-risk when their current charge is especially serious.[40] Thus, in each of these cases, judges seem to assign too much value to the current offense in comparison to other relevant factors including the defendant's prior criminal record,

33 Sunstein (2019), p. 2.

34 Kleinberg et al (2018), p. 241.

35 Even though I maintain that these differences are still being very much exaggerated; for an insightful analysis on this topic, *see* Pejovic (2001).

36 Baughman (2017), p. 15.

37 Sunstein (2019), p. 501.

38 Of course, not all identified biases occur globally in the same way. However, as we shall see, chances are that the bias in question will occur not only in the US. See also Dhami and Ayton (2001) showing that human adjudicators in the UK follow simple heuristics.

39 Sunstein (2019), p. 502.

40 *Ib*.

age, and employment history.[41] Sunstein refers to this phenomenon as the *current offense bias*.[42]

He then links the current offense bias to the well-known *availability bias*, i.e., the tendency to overestimate the likelihood of an event occurring in the future if examples can easily be brought to mind.[43] Even though one may be a bit skeptical as to the degree of relatedness between the current offense bias and the availability bias,[44] the overall point stands: judges in different jurisdictions are likely to make the same mistakes because there is no reason to assume that Asian, African, or European judges will not suffer from the current offense biases. This is especially so because the related availability bias seems to be a general trait of the mind trying to access the probability of an event occurring based on associative distance.[45]

We have now seen that crime and detention rates can be significantly decreased without perpetuating or even increasing racial discrimination. This itself could be viewed as a welcome development, as the absolute numbers of minorities in jail would decrease. But does that mean that, if we adopt Kleinberg and colleagues' algorithm, the percentage of African Americans and other minorities in prison will remain the same? If we assume that human decision-making will eventually decrease discrimination, and thereby reduce not only the absolute but also the relative number of African Americans in prison, the question arises whether algorithms would lag behind HI after some time has passed. It seems plausible to assume that both absolute and relative detention rates for minorities will decrease in the future, if a human-centered judiciary follows humanity's track record on moral progress.[46] However, it will take time to develop the necessary and crucial cognitive features to do so more fully, such as through enhancement of the abilities to override implicit biases.[47] At the same time, it is possible to instruct an algorithm in a way

41 *Ib*.

42 *Ib*.

43 Tversky and Kahneman (1982).

44 Sunstein (2019), p. 502 views them as "close cousins".

45 Tversky and Kahneman (1973). This having been said, the cross-jurisdictional application of the current offense bias is ultimately an empirical issue that requires empirical confirmation.

46 Pinker (2011; 2018).

47 I leave aside other notable problems these approaches might bring about, such as concerns related to privacy and freedom of thought. I also leave aside the question to what degree discriminatory outcomes within the US criminal justice system are currently influenced by implicit racial bias, or whether other factors, such as education, poverty, and higher police presence in African American communities are the main drivers (while acknowledging that factors such as these are in turn influenced by

that it produces morally or socially desirable outcomes, whatever these may be. For instance, in one scenario, Kleinberg and colleagues instructed the algorithm to maintain the same detention rate while equalizing the release rate for all races. Given this, the algorithm was still able to reduce the failure to appear and re-arrest ("crime") rate by 23%.[48] In another simulation, the algorithm was instructed to produce the same crime rate that judges currently achieve. Under these conditions, a staggering 40.8% fewer African Americans and 44.6% fewer Hispanics would have been jailed.[49]

We might wonder whether these results really indicate that human decision-making *ought* to be *supplemented* or even (partially) *replaced* by AI from the perspective of discrimination. If it is all about cognitive biases – whether they are related to implicit racial biases or about current offense bias – maybe we can try to improve human decision-making before we get rid of it entirely. Indeed, it is plausible to mitigate some biases to some degree. For instance, the strength of the availability bias can even be reduced by simply being aware of it.[50] Additionally, I have previously suggested different institutional and procedural changes as well as mandatory training in behavioral economics and cognitive biases for members of the judiciary in order to mitigate biases.[51] However, it does not seem feasible to eliminate them altogether this way.[52] To put it simply, in order to completely get rid of cognitive biases at this day and age, one must get rid of human decision-making.[53]

To summarize, Kleinberg and colleagues' findings enormously strengthen the case for AI in the judiciary from the perspective of decreasing discrimination. This is especially so because it is possible to reduce crime and detention rates at the same time. Although there is much more to say about human decision-making, cognitive

structural discrimination). For a critical note on the role of implicit bias, see, for instance, Oswald et al. (2013).

48 Kleinberg et al. (2018).

49 Kleinberg et al. (2018).

50 Gigerenzer (1991). For more on countering implicit biases, *see* Teal et al (2012), Boscardin (2015), Ingriselli (2015), Shaked-Schroer (2008).

51 Winter (2020).

52 *Ib*.

53 *Ib*. Note, however, that this does not imply that an AI-based judiciary does not face *any* of such risks. Of course, the design of the AI itself as well as the selection of training data depend once again on human decision-making. This said, it may be more realistic to avoid biases within said design and selection process compared to debiasing a human-centered judiciary as indicated by the results of Kleinberg and colleagues' study.

biases, and the role of empathy therein,[54] contrary to the present popular belief, AI *can* be a "force for (…) equity."[55] While it is important to caution that Kleinberg and colleagues' algorithm was merely concerned with bail decisions, it seems reasonable to assume that it would generally be easier to prove whether direct discrimination occurred under algorithmic than human judicial decision-making. In fact, if the decision procedure is transparent, it would be so obvious that it is unlikely to occur in the first place. The preferred trade-offs among crime, detention, and discrimination would still have to be made, but regardless of how one decides, the result would be an improvement from HI. Having said this, it is important to note – as experienced in the COMPAS case – that the exact trade-offs made may not always be visible. So, should one reject the implementation of AI, even if it can reduce discrimination, for lacking transparency?

II. FROM TRANSPARENCY OF PROCEDURES TO TRANSPARENCY OF OPTIONS

The relationship between democracy and transparency is highly contested. Contrary to widespread intuition, it may even be the case that authoritarianism leads to more transparency than democracy. This may be so because greater vulnerability to public disapproval could make democratically elected officials more inclined to promote opacity or withhold information compared to their autocratic counterparts, who have to worry less about public perception.[56] The fact that democratic governments have incentives to obfuscate their policies can be illustrated by the recent remarks of Germany's Minister for the Interior Horst Seehofer, who recently stated that "you have to make laws complicated".[57] Despite the fact that the statement itself may arguably not have been well thought-through from a political standpoint, and Seehofer consequently shortly thereafter tried to argue that he was being "slightly ironic", the statement does get at the heart of the tension between democracy and transparency. If passing a specific law is *better* for citizens than any of the alternatives, from the perspective of elected officials, but the law in question is very unpopular, should it still be adopted? And if so, is it permissible to communicate it

54 Empathy is often raised as an argument against AI. But note that it may not always be the compassionate, nice concept we often take it to be and may even be responsible for much of discrimination. *See,* in general, Bloom (2016).

55 Kleinberg et al (2018) p. 241.

56 Hollyer et al. (2011).

57 For more information regarding the context of his remarks, *see* Das Gupta and Fried (2019).

in a way that makes re-election more probable, despite the fact that citizens disagree with the measure being taken, even if they ultimately benefit from the new law?

The argument that democracy may *in fact* not always be the ideal engine for transparency does not entail that transparency is not a crucial factor thereof from a *normative* perspective. Of course, transparency broadly understood helps to fight corruption, promote trust in public institutions, and contribute to the public discourse.[58] Yet, in order to answer more specific questions regarding the state of transparency when it comes to artificial judicial decision-making from a democratic perspective, one first needs to consider the function of transparency within a democracy. Interestingly, despite the visibility of transparency as a concept in the public discourse, surprisingly little attention has been paid to its underlying purpose.[59] Gupta even calls it an "overused but under-analyzed concept".[60] The basic argument for transparency being an essential part of *any* democratic system runs along the following lines:[61]

1. Democracy requires citizens to take an active part in politics.
2. For citizens to be able to participate in politics, they need to be able to make political judgments based on relevant information.
3. In order to make political judgments based on relevant information, citizens need to have access to that relevant information.
4. In order for citizens to have access to relevant information, such information must be transparent.

The central question in this framework is what exactly counts as *relevant* information. Do citizens need to know *who* made the decision in question and *how*, or do they also need to be aware of alternative routes which could have been taken? What information should be available to citizens when it comes to decisions by the judiciary, rather than the executive or legislature?

---

58 Note, however, that some research points in the direction that high levels of transparency may also have negative consequences. See, among others, Fox (2007), Licht (2011), Licht (2013) and Moore (2018). The scope of this paper forces me to solely focus on the liberal democratic perspective of transparency, and I will leave aside other theories, perspectives, and approaches to the matter which may or may not shift one's opinion, particularly on the desired degree of transparency and noteworthy exceptions.

59 Moore (2018).

60 Gupta (2008).

61 *See*, among others, Dahl (1971), who argues that any conception of democracy requires the free flow of information in order to make informed choices, compared to a minimalist approach to democracy as taken by Schumpeter (1942) or, more recently, Przeworski et al. (2000), who define democracy as a regime in which the executive and the legislature are both filled by "contested elections".

If, as mentioned above, democracy requires citizens to take an active part in politics, then relevant information could be any information helpful for making political choices such as who or what to vote for, what to protest, and when and how to engage in public discourse more generally. As Bellver and Kaufmann argue, the information provided needs to account for the performance of public institutions because transparency is a tool to facilitate the evaluation of public institutions.[62] More importantly, not only the decisions made by directly or indirectly elected individuals, but state actions more generally, including decisions and judgments made by the judiciary,[63] need to be transparent in order for citizens to be sufficiently informed.

In order to evaluate public institutions carefully, citizens ideally need to know (or at least be able to know) not only the outcome of the decision itself, the people involved, and their respective roles in the process but – importantly – also alternative options which could have been taken. If, for instance, a decision is not very popular, but the alternatives are significantly worse, then ideally this would be communicated just as it should be transparent if a decision was made which intuitively sounds good, but which arguably had much better alternatives from the perspective of citizens. Accordingly, in order to guarantee that citizens receive relevant information to evaluate public institutions, they should be able to learn what kind of decision was made, the procedures leading to it, and what other options were available.[64] To put it simply, *transparency about options* may be just as important as transparency about procedures and actors from the liberal democratic point of view.

Since it would be an improvement from a liberal democratic point of view if options were more frequently transparent, some AI – as illustrated by the clear trade-offs Kleinberg and colleagues' recidivism algorithm produces – has strong advantages over HI in this regard. While nowadays citizens and policymakers alike have to trust their (often unreliable) intuitions to assess the situation, the clarity of the trade-offs, for example between public safety, discrimination, and detention rates, as in the case of bail decisions, will become clearer with the implementation of AI. For instance, this means that citizens will not only be able to see the racial composition of the population denied bail, but will also know the consequences of lowering or increasing this rate. In sum, citizens will have better access to highly

---

62 Bellver and Kaufmann (2005, p. 5); *cf.* also Florini (1999, 5).

63 As noted by Liptak (2008), the United States is the only country that elects a *significant* portion of its judges directly.

64 *Cf.* also Licht and Licht (2020), who distinguish among (a) transparency that informs about final decisions or policies, (b) transparency with regards to the process resulting in the decisions, and (c) transparency about the reasons on which the decision is based.

relevant information in order to participate in the political process.[65] Indeed, for better or worse, it will be much more difficult for politicians like Seehofer to use complicated laws to obscure the political decision and evaluation process. This is the case even if programs do not use explicitly programmed decision-making (e.g., a simple decision tree or a more complex algorithm), but instead use machine learning, including deep learning, due to the fact that the relevant trade-offs can still be transparent by evaluating actual and simulated impact, regardless of the complexity of machine learning.

Yet most commentators in the debate on AI and transparency take a different stance and focus. Instead of concentrating on the new levels of transparency regarding the tradeoff being made, they rightly point out that the algorithm itself is often unknown.[66] Case in point, the COMPAS risk assessment algorithm, which is now used for bail and sentencing decisions in more than 20 jurisdictions within the United States, is a protected trade secret and thus remains a black box.[67] This is already problematic because without knowing the algorithm or machine learning training set, one cannot understand and challenge the decision, in case the design is flawed. From a liberal democratic point of view, this is even more troublesome. If citizens are not able to obtain vital information to evaluate the operations of the judiciary and related procedural norms, they cannot participate in the democratic process in a meaningful way, aside from pointing out that the algorithm should be made public. It may not be unethical *per se* to outsource typical tasks of the judiciary to private companies, yet it becomes highly questionable if this process leads to the lack of transparency witnessed in the COMPAS case.

Hence, while AI has strong potential benefits with regards to the clarity of tradeoffs, the current lack of transparency (of procedures) in many cases and the potential lack thereof in the future has led scholars to become increasingly skeptical.[68] However, in order for it to be sufficient reason to reject the implementation of AI, one would have to show that, at first, HI is in fact better suited to tackle these issues and, secondly, the lack of this kind of transparency is more

---

65 Sunstein (2019), p.7 even argues that the clarity of tradeoffs "may be the most important point" with regards to the recidivism algorithm.

66 In this vein, among others, Floridi et al. (2018); O'Neil (2016); Wachter et al. (2017).

67 Piovesan and Ntiri (2018).

68 For instance, the number one recommendation of AI Now Institute's annual report states that "core public agencies, such as those responsible for criminal justice, healthcare, welfare, and education (i.e., "high stakes" domains) should no longer use 'black box' AI and algorithmic systems." AI Now (2017), p. 1. The report emphasizes that 'black box' systems are especially vulnerable to skewed training data and the replication of human biases due to a lack of transparency (p. 15). *Cf.* also Završnik (2020).

harmful from the liberal democratic perspective than the lack of transparency regarding the above-mentioned trade-offs. I will consider these in turn.

It may be argued that human judicial decision-making is much more transparent than artificial judicial decision-making because citizens can identify human judges. They can read their arguments and decide for themselves whether the judgment was reasonable or not. In short: Human judges *explain* why they have come to the decision they came to. Although one can find plenty of ongoing research with regards to the explainability of AI, present applications certainly lag behind human capabilities to explain.[69] However, there are notable problems with this line of reasoning.

First, giving human explanations does not equal full transparency of procedures. Explaining judgments is important and can provide a basis for appeal, but an explanation does not necessarily reveal the true underlying cognitive processes. The processes of human (judicial) decision-making are just as unknown as the proceedings of some applications of AI. Following a computational theory of cognition,[70] one might even argue that we can at least know the algorithm or data set with regards to artificial decision-making, even if we may not understand its operations, whereas the algorithm running human decision-making remains unknown. It is crucial to note that while a computational perspective of cognition certainly strengthens this argument, it does not rest upon it. Whatever philosophy of mind one might prefer, human decision-making remains a black box as well.

Second, human explanations are not only prone to error and cognitive biases, but also at risk of rationalizing underlying motivations and preferences.[71] Intuitively, one might expect that this general tendency might be mitigated by expert knowledge and education, yet as studies have shown over and over again, experts rarely perform better than lay-people.[72] Accordingly, the worry is that even though one might know and understand the official reasons given, they might be misleading and the risk remains that the ultimately decisive reasons are still unknown. This is not only important for democratic evaluations but might also decrease the chances of a potential successful appeal.

---

69 *Cf. See* Waltl and Vogl (2018), Barredo Arrieta et al. (2020); the technical challenges of explainable AI (XAI) are also referred to as the *interpretability problem*.

70 Piccinini (2016), Chalmers (2011).

71 Cushman (2020) with further references.

72 *Cf.* the comprehensive list covering research on military leaders, engineers, accountants, doctors, real estate appraisers, option traders, psychologists, and lawyers presented in Guthrie et al. (2001); see also Meadow and Sunstein (2001). See generally Kahneman and Tversky (1983). Research specifically focused on judicial decision-making can be found in Englich et al. (2006); Wistrich et al. (2015); Rachlinski and Wistrich (2017); Wistrich and Rachlinski (2018); Struchiner et al. (2020); Winter (2020).

Third, depending on the complexity of the individual judgment, one might question whether laypeople can fully understand judgments anyway. Surely, one might counter-argue that lawyers will be able to understand the judgment and advise their client accordingly. However, from a liberal democratic perspective, there is not much benefit in lawyers navigating cases if the vast majority of citizens are unable to comprehend and – based on this – evaluate judicial norms and behavior correctly. Arguably, overconfidence may even lead laypeople to think they understand the judgment and give them a false sense of understanding.

These theoretical and practical concerns question the assumption that HI really does provide more liberal-democratically relevant transparency than AI. However, let us assume that this analysis is either mistaken or that there are other overriding arguments which speak in favor of HI-transparency in this regard.[73] Consequently, one would still have to investigate whether the lack of human explanations for individual judgments is more harmful than the lack of transparency regarding the above-mentioned trade-offs. Should humans have the *possibility* to understand judgments in individual cases, or should they rather be able to clearly see the necessary trade-offs behind rules, which not only shape society but also the very individual judgments they care to understand? Should they understand the purpose of the rules, its underlying trade-offs and the explanations given by the ones they voted for (legislature) or the application of them made by often not democratically accountable judges in selected scenarios? Even though the advantages of having access to human judicial explanations ought not to be downplayed, I am afraid that the transparency of the trade-offs in question might simply be even more important.[74]

III. FROM SEPARATION OF POWERS TO JUDICIAL DEPENDENCE

While improvements with regards to discrimination (I.) and transparency (II.) seem possible, and related issues have been receiving the appropriate academic attention in the past few years, medium- and long-term threats arising from the application of advanced AI within the judiciary for liberal democracy have been almost completely neglected.[75] Hence, this part of the analysis aims to raise awareness of these risks. More precisely, I will make the point that advanced artificial judicial intelligence

---

73 For instance, one might argue that AI reasoning is sufficiently alien to human cognition that humans will always (think that they) understand the thought processes of another HI better than the operations of an AI.

74 Assuming that human judicial explanations are more important than transparent trade-offs, one would still have to show that this also justifies higher incarceration rates – both generally and especially regarding minorities.

75 One exception which will be discussed below is the work by Michaels (2020).

may threaten judicial independence and the separation of powers more generally. Before I proceed with the analysis by outlining why the development of artificial judicial intelligence ought to be followed closely from the point of view of liberal democracy, it is necessary to provide a brief explanation for this seemingly science-fiction scenario.

1. The Possibility of an Advanced Artificial Judicial Intelligence

Advanced artificial judicial intelligence (AAJI) can be defined as an artificially intelligent system that matches or surpasses human decision-making in all domains relevant to judicial decision-making. Crucially, this does not require the development of an artificial general intelligence (AGI) while still avoiding the need for hybrid human-AI judicial systems. Such a system goes significantly beyond the current state of the art; with it, humans would have the ability to outsource decision-making within the judicial sector entirely if they wished to do so. Although such a development is often considered by legal researchers to be extremely unlikely in the medium- or even long-term future,[76] machine learning researchers think that even the development of an AGI, i.e., an artificially intelligent system that matches or surpasses human decision-making in *all* relevant domains,[77] is much closer than the common sense among jurists seems to indicate.[78] For instance, experts believe there is a 50% chance of AI outperforming humans in all tasks in 45 years.[79] With regards to specific activities, they predict that AI will outperform humans fairly soon. This includes translating languages (by 2024), writing high-school essays (by 2026), driving a truck (by 2027), working in retail (by 2031), writing a bestselling book (by 2049), and working as a surgeon (by 2053).[80]

Surely, one can debate whether it is even possible for an AI to engage in legal reasoning. Crootof, for instance, points out that "the judgment we value in a common law process is a distinctively human skill."[81] In this regard, she further argues that "[g]iven their sensitivity to changing social norms, human judges are uniquely able to oversee legal evolution and ensure that our judicial system 'keep[s] pace with the times.'" However, if one is primarily concerned about detecting (changing) social

---

76 See Martinez and Winter (2021).

77 On the notion of AGI, see Wang and Goertzel (2006). Other terminologies being used to describe the same or a very similar technologies are "strong AI", "human-level AI", "true synthetic intelligence" and "general intelligent system".

78 For expert surveys on AGI timelines, see Baum et al. (2011); Müller and Bostrom (2016); Grace et al. (2018); Gruetzemacher et al. (2019).

79 Grace et al. (2018).

80 Grace et al. (2018).

81 Crootof (2019), p. 238, citing Kaminski (2019).

norms, AI may arguably be better equipped to do so than HI. While AI can be trained with vast data sets from diverse sources and be more easily and broadly adapted with changing norms, HI relies on personal interpretations of comparatively little and highly selected information from media and interactions with again a very selected group of people. Hence, AI not only has the power to process a greater amount of relevant data, but it is also less likely to fall for the ubiquitous confirmation bias and other misleading selection mechanisms that lead one to paint an inaccurate picture of existing social norms.[82]

This having been said, following a Dworkinian interpretation of law, one could certainly question whether even an advanced AI would be up for the task of "moral reasoning" as part of the judicial decision-making process.[83] At the same time, chances that the judicial system can be outsourced would arguably increase if one would follow Oliver Wendell Holmes' prediction theory of law.[84] Holmes famously stated that "[t]he prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law."[85] The purpose of this section is not to convince the reader of the merits of a specific legal theory, but rather to emphasize the high degree of jurisprudential uncertainty.[86] If we can agree that AI would be able to take over the judiciary from *some* legal theoretical perspectives – even if one personally does not share those views or considers them unlikely to be correct – we have strong reason to start thinking about the accompanying consequences and potentially relevant safeguards. In other words, the mere possibility of an AAJI should lead us to take seriously its implications.

---

82 Again, assuming that it is easier to avoid biases within the AI design and data selection process compared to debiasing a human-centered judiciary. *See also supra* mote 53.

83 Note, however, that even this is far from clear. For instance, one might argue that Dworkin's (2011) description of moral reasoning as "the interpretation of moral concepts" (p. 102) would not necessarily exclude advanced AI, which would presumably be as or even more capable than humans at the "integration of background values and concrete interpretive insights" (p. 135).

84 *See* Holmes (1897), p. 457, 461.

85 Holmes (1897), p. 461.

86 Discussions on "jurisprudential uncertainty" understood as "normative uncertainty with respect to legal theory" (Winter et al., 2021, p. 97) have only just started. *See* Winter et al. (2021); Winter (2021). The related debate in ethics has been getting more attention recently. See, among others, Gustafsson and Torpman (2014); Lockhart (2000); MacAskill (2014); MacAskill et al. (2020), Tarsney (2018). *See also* Barry and Tomlin (2019) who apply moral uncertainty to different issues in criminal law theory, including sentencing and criminalization theories.

2. Advanced AI & Judicial Independence

Even though one might argue that a liberal democratic system may in theory be achieved and sustained without the separation of powers, this separation of powers has come to be a cornerstone of any liberal democratic system around the world. The separation of the judicial, legislative, and executive branches of government serves a vital purpose in minimizing the concentration of power and maintaining checks and balances across the government. While checks and balances differ in their configuration across jurisdictions, any significant influence on a given power structure (e.g., branch of government) will destabilize the system. Relatedly, this destabilization can take very different forms and outcomes in different systems. For instance, parliamentary democracies in Western Europe may respond very differently to the implementation of an AAJI in comparison to the presidential systems in the Americas. Additionally, those jurisdictions and cultures which consider their judicial branch to have some legislative-like lawmaking powers, such as the European Union[87] and the United States,[88] may evaluate the threats and opportunities imposed by AAJI very differently than those who consider their judiciary as primarily politically neutral (even if this may in fact not be the case, as legal realism tells us), such as the United Kingdom[89] or Germany.[90] This fact can be illustrated by the intensity of discussions taking place during the selection procedure of Supreme Court Justices. While appointments in the United States can occupy media attention for weeks,[91] there is comparatively little debate in Germany or the UK, which is unsurprising if the judiciary is considered to apply laws neutrally. Finally, the crucial question is not *whether* but *how* the interaction and power dynamics between branches of government may be threatened by an AAJI. Will the judiciary become more or less powerful? How would this affect the interaction with

---

87 Barnard (2013), p. 377.

88 *Cf.* Michaels (2020), p. 1098 argument in this regard: "Courts exercise an important lawmaking and policymaking function when they interpret the law so as to resolve legal questions, and it is beneficial for such interpretation to take place in the context of concrete factual disputes."

89 *See*, for instance, the still relevant observation by Denning (1963): "According to the British conception it is vital that the judges should be outside the realm of political controversy. They must interpret the law and mould it to meet the needs of the times, but they cannot bring about any major alterations in policy. It is only thus that judges can keep outside the sphere of politics" (p. 300).

90 Kischel (2013) analyzes the election procedure of German constitutional court justices from a comparative perspective and finds that it is "necessary for its proper □politically□ neutral functioning."

91 For an overview of the Kavanaugh hearings in 2018, *see* Grynbaum (2018), Bowden (2018).

and powers of the legislature and executive? Such questions are important from any liberal democratic perspective because, first of all, authoritarianism tends to favor a weak judiciary. Second, an independent judiciary is an important ally in the fight for minority rights and preventing the "tyranny of the majority".

Michaels argues that the current human-based judiciary is necessary to maintain the separation of powers, for only a human-centered judiciary raises ample attention to the law from the involvement of judges, lawyers, law professors, and so on.[92] For Michaels, it seems likely that the legal community would diminish if artificial judicial intelligence were to replace human judges,[93] and without a legal community, humans would pay little attention to the law.[94] He further argues that such lack of sufficient attention from the legal community might have devastating effects. First, it would be hard to imagine any public response to abuses of authority or concentration of power across the three branches.[95] Second, there would generally be "little incentive to construct high-quality legal arguments if there was no possibility that doing so could shape the result."[96]

However, even assuming that the legal community would in fact diminish or be significantly reduced, it is not clear that this *itself* would necessarily lead to great power imbalances due to decreased human attention. On the contrary, there may be some significant accompanying advantages resulting from the shift in focus. The attention nowadays spent on specific cases may shift to law-making processes and the analysis of the overall consequences of the laws in question. While society is often occupied with extraordinary individual cases of little systemic value for the long-term, with AAJI, humans would be able to focus on the evaluation of abstract principles and rules rather than on our intuitive judgments about specific, often highly politicized, cases. Given that such perceptual intuitions are especially prone to cognitive biases,[97] offer little normative guidance, and have limited value for the whole of society, this may even nudge the public discourse onto a more beneficial path. Instead of focusing on extraordinary individual cases, attention might shift to the much more important trade-offs between detention, crime rates, and

92 Michaels (2020), p. 1096

93 Michaels (2020), p. 1096; *cf.* also Volokh (2019); Re and Solow-Niederman (2019).

94 Michaels (2020), p. 1096. It should be noted that Michaels does not specify how advanced the AI would have to be for such a scenario to occur. However, it is clear that he has a less capable AI in mind than what has been defined as an AAJI in this article. Presumably, his argument would apply all the more in the case of AAJI.

95 Michaels (2020), p. 1096.

96 Michaels (2020), p. 1097, *see* also p. 1084. Although this point is not crucial for the argument that follows, it may well be possible to have adversarial AIs arguing on behalf of each party, with an AAJI adjudicating.

97 The terminology "perceptual intuition" goes back to Sidgwick (1907).

discrimination discussed in the first part of this article – potentially leading to lower incarceration, higher security, and less discrimination. Furthermore, the shift in focus from the judiciary onto the legislature (and executive) may carry the advantage of greater public accountability of the legislature, which would again profit from the transparency of the aforementioned trade-offs.

Although one can debate whether the adoption of AAJI would in fact lead to decreased attention on the judiciary, and whether such development would necessarily go hand in hand with a loss of power, the judiciary may nonetheless lose a significant share of its power and, eventually, its independence. This is ultimately so because taking the human out of the equation faces risks of democratic legitimacy. This itself may lead to a reduction of judicial power. Additionally, with reduced democratic legitimacy, it is difficult to imagine that judicial discretion would remain a decisive factor in many cases. On the contrary, I assume that legislatures would use a democratically weak judiciary to justify more and more specific rulings, leaving little discretion for the AAJI. With power shifting from the judiciary to the legislature, judicial independence and the separation of powers more generally would be under threat. The graphic below outlines the threat to judicial independence resulting from the implementation of AAJI.

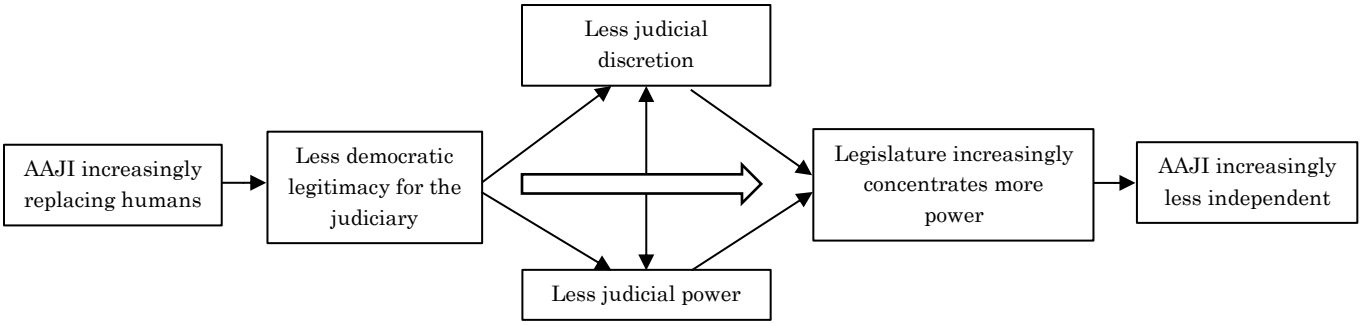

Figure 1*. AAJI, legitimacy, and judicial independence.*

If democratic legitimacy were compromised in this way, the people might even welcome more powerful legislatures (and executives). Such a development could be accelerated by the improved accountability of the legislature resulting from greater transparency of the aforementioned trade-offs. At last, the legislature may increasingly be perceived as responsible for the courts' rulings. Case in point, if the AAJI does not decide as the public wishes, people may blame the legislature for – in their eyes – a poor design of the laws in question, or poor decision to use a biased AAJI, rather than blaming the AAJI for poor implementation or interpretation of such laws. From this perspective, even governing parties who strongly support judicial independence might give in to public pressure to limit AAJI's discretion over time.

This is not a definite outcome, and it is highly uncertain where this shift in power would ultimately lead. However, the adoption of an AAJI may pose a great risk for judicial independence which ought not to be taken before necessary safety mechanisms will have been generated. While there are tremendous efforts invested into the development of more advanced AI, including AGI and AAJI, little attention has been paid to creating the necessary political infrastructure that would allow society to harness the potentially enormous benefits of AI without risking the collapse of judicial independence. Although I disagree with Michaels regarding the reason why one might expect a shift in power (attention vs. legitimacy), this analysis strengthens the case for AAJI affecting the separation of powers. More precisely, I argue that it poses a great threat to judicial independence, which ought not be taken lightly.

IV. CONCLUSION

As part of the legal community, we might be inclined to improve the judicial system as much as possible and defend it from potential threats, even if those threats accompany extraordinary improvements in other domains. As academics, we want to completely understand AI reasoning before we recommend its implementation, even if we do not fully understand human reasoning either. And as humans, we tend to prefer the status quo, even when change would be net-positive.[98] All these motivations might explain why so many of us have a strong aversion to the implementation of AI into the judicial sector, but, unfortunately, motivations are not ideal truth-tracking processes.[99] At least when it comes to issues of discrimination and transparency, this analysis makes clear that current applications could come with great advantages. Whereas this proposition can be defended from a wide range of normative perspectives on discrimination, the improvements with regards to transparency depend on how or whether one favors transparency of options over transparency of procedures and agents. Having said this, one should bear in mind that the existence of the possibility of such improvements does not mean that there is no risk of large-scale discrimination or a lack of transparency. As seen, the fact that such AI is *possible* does not mean that AI used *in practice* will be more transparent or less discriminatory.

While the analysis hitherto focused on existing technological capabilities, the final chapter introduced the concept of AAJI, defined as an artificially intelligent

---

98 *Cf.* Samuelson and Zeckhauser's (1988) analysis of the status quo bias. Eidelman and Crandall (2012) further note that the status quo bias creates "barriers to cognitive and social change" (p. 270).

99 On the contrary, given the intuitive appeal of these motivations, one might consider debunking arguments.

system that matches or surpasses human decision-making in all domains relevant to judicial decision-making. Concentrating on the potential power shift from the judiciary to the legislature, it has been argued that the adoption of an AAJI threatens judicial independence – and with that one of the central foundations of liberal democracy.

Needless to say, this attempt to outline the challenges of artificial judicial decision-making for liberal democracy is far from complete, and further research is needed in many regards. First, it is necessary to identify what other challenges the adoption of an AAJI may bring about, such as the risk that algorithms may grant laws a new kind of permanency and thereby create an additional barrier to legal evolution.[100] Second, it will be crucial to investigate how one can uphold liberal democratic values more generally and, in particular, the separation of powers with an AI Judiciary. The short-term benefits AI could offer over HI with regards to access to justice, transparency, and fairness are so enormous that one may overlook the long-term threats imposed by AAJI. Working out the political and legal infrastructure to reap the fruits of artificial judicial intelligence in a safe and stable manner should become a priority of future research.

[100] Crootof (2020) refers to this phenomenon as "technological-legal lock-in".

BIBLIOGRAPHY

AI Now (2017) AI Now 2017 Report. New York University, New York. ainowinstitute.org/AI_Now_2017_Report.pdf

Aletras et al. (2016) Predicting Judicial Decisions of the European Court of Human Rights: a Natural Language Processing Perspective. PeerJ Computer Science. https://doi.org/ 10.7717/peerj-cs.93

Angwin J et al. (2016) Machine Bias. ProPublica. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 25 June 2020

Armstrong S, Sotala K, hÉigeartaigh SSÓ (2014) The errors, insights and lessons of famous AI predictions – and what they mean for the future. Journal of Experimental & Theoretical Artificial Intelligence, 26, 317–342. https://doi.org/https://doi.org/10.1080/0952813X.2014.895105

Barnard C (2019) The Substantive Law of the EU, 6th Edition. Oxford University Press, Oxford, New York

Barredo Arrieta A et al. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barry C, Tomlin P (2019) Moral uncertainty and the criminal law. In: Alexander L, Ferzan KK (eds) Palgrave Handbook of Applied Ethics and the Criminal Law. Palgrave Macmillan/ Springer Nature, Cham, Switzerland

Baum S (2018) Reconciliation between Factions Focused on Near-Term and Long-Term Artificial Intelligence. AI & Society 33(4):565–572

Baum S (2020) Medium-Term Artificial Intelligence and Society. Information 2020, 11(6):290–305. https://doi.org/10.3390/info11060290

Baum SD, Goertzel B, Goertzel TG (2011) How long until human-level AI? Results from an expert assessment. Technological Forecasting and Social Change 78:185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Bloom P (2016) Against Empathy: The Case for Rational Compassion. Ecco Press, New York

Boscardin C (2015) Reducing Implicit Bias Through Curricular Interventions. Journal of General Internal Medicine 30(12):1726–1728

Bostrom N (2006) How Long Before Superintelligence. Linguistic and Philosophical Investigations 5:11–30

Bostrom N (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford

Bostrom N, Ord T (2006). The reversal test: eliminating status quo bias in applied ethics. Ethics 116(4):656–679. https://doi.org/ 10.1086/505233.

Bowden J (2018) Timeline: Brett Kavanaugh's nomination to the Supreme Court. The Hill. https://thehill.com/homenews/senate/410217-timeline-brett-kavanaughs-nomination-to-the-supreme-court. Accessed 26 June 2020

Carneiro D et al. (2015) Online Dispute Resolution: An Artificial Intelligence Perspective. Artificial Intelligence Review 41:227–228

Carnes M et al. (2015) The effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. Academic Medicine 90(2):221-230

Cave S, Ó hÉigeartaigh S (2019) Bridging near- and long-term concerns about AI. Nature Machine Intelligence 1:5–6

Chalmers D (2011) A Computational Foundation for the Study of Cognition. The Journal of Cognitive Science 12(4):325-359. https://doi.org/10.17791/jcs.2011.12.4.325

Chander A (2017) The Racist Algorithm? Michigan Law Review 115(6):1023-1045

Chen D (2019) Machine Learning and the Rule of Law. In: Livermore M, Rockmore D (eds) Law as Data. Santa Fe Institute Press, Santa Fe, pp 433–441

Chohlas-Wood A (2020) Understanding risk assessment instruments in criminal justice. Brookings. www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice. Accessed 26 June 2020

Crootof R (2019) "Cyborg Justice" and the Risk of Technological-Legal Lock-In. Columbia Law Review Forum 119:233–251

Cushman F (2020) Rationalization is rational. Behavioral and Brain Sciences 43. Cambridge University Press (43): E28. https://doi.org/10.1017/S0140525X19001730

Dabney D et al. (2017) American Bail and the Tinting of Criminal Justice. The Harvard Journal of Crime and Justice 56(4):397–418

Dahl R (1971) Polyarchy; participation and opposition. Yale University Press, New Haven

Das Gupta O, Fried N (2019) Seehofer redet über Gesetzestrick - hinterher spricht er von Ironie. Süddeutsche Zeitung. www.sueddeutsche.de/politik/seehofer-datenaustauschgesetz-1.4479069. Accessed 26 June 2020

Denning L (1963) The Function of the Judiciary in a Modern Democracy. Pakistan Horizon 16(4):299–305

Devine P et al. (2012) Long-term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention. Journal of Experimental Social Psychology 48(6):1267–1278

Dhami MK, Ayton P (2001), Bailing and jailing the fast and frugal way. Journal of Behavioral Decision Making 14:141–168. https://doi.org/10.1002/bdm.371

Dworkin R (2011) Justice for Hedgehogs. Harvard University Press, Cambridge

Eidelman S, Crandall C (2012) Bias in Favor of the Status Quo. Social and Personality Psychology Compass 6(3):270–281. https://doi.org/10.1111/j.1751-9004.2012.00427.x

Englich B et al. (2006) Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. Personality and Social Psychology Bulletin 32(2):188–200. https://doi.org/10.1177/0146167205282152

Floridi L, Cowls J, Beltrametti M et al. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Florini A (1999) Does the Invisible Hand Need a Transparent Glove? The Politics of Transparency. Paper presented at Annual World Bank Conference on Development Economics, Washington, D.C., 28–30

Fox J (2007) Government Transparency and Policymaking. Public Choice 131(1/2):23–44.

Gigerenzer G (1991) How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." European Review of Social Psychology 2:83–115

Goertzel B, Pennachin C (Eds) (2007) Artificial General Intelligence. Springer-Verlag, Berlin, Heidelberg

Grace K, Salvatier J, Dafoe A, et al (2018) When Will AI Exceed Human Performance? Evidence from AI Experts. Journal of Artificial Intelligence Research 62:29–754

Gruetzemacher R, Paradice D, Lee KB (2019) Forecasting Transformative AI: An Expert Survey. arXiv:190108579 [cs]

Grynbaum MM (2018) Kavanaugh Hearings on TV Offer Riveting Drama to a Captive Nation. The New York Times. https://www.nytimes.com/2018/09/27/business/media/kavanaugh-blasey-ford-hearing-tv.html. Accessed 26 June 2020.

Gupta A (2008) Transparency Under Scrutiny: Information Disclosure in Global Environmental Governance. Global Environmental Politics 8(2):1–7

Gustafsson JE, Torpman O (2014) In Defence of My Favourite Theory. Pacific Philosophical Quarterly 95:159–174. https://doi.org/10.1111/papq.12022

Guthrie C et al. (2001) Inside the Judicial Mind. Cornell Law Review 86(4):777–830

Hacker P (2018) Teaching Fairness to Artificial Intelligence; Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law. Common Market Law Review 55:1143–1185

Hammer P (2018) Detroit 1967 and Today: Spatial Racism and Ongoing Cycles of Oppression. Journal of Law in Society 18(2):227–235

Hollyer et al. (2011) Democracy and Transparency. The Journal of Politics 73(4):1191–1205. https://doi.org/ 10.1017/s0022381611000880

Holmes, O (1897) The Path of the Law. Harvard Law Review, 10:457–ADD END of page range

Huq AZ (2020) A Right to a Human Decision. Virginia Law Review 106:611–688

Ingriselli E (2015) Mitigating Jurors' Racial Biases: The Effects of Content and Timing of Jury Instruction. Yale Law Journal 124(5):1690–1745

Kahneman D, Tversky A (1982) 30 - Intuitive prediction: Biases and corrective procedures. In: Kahneman D, Slovic P, Tversky A (eds) Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, pp 414–421

Kahneman D, Tversky A (1983) Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. Psychological Review 90(4):293–315. https://doi.org/10.1037/0033-295X.90.4.293

Kaminski M (2019) Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability. Southern California Law Review 92:1529–1616

Karnow C (2008) Setting Bail for Public Safety. Berkeley Journal of Criminal Law 13(1):1–30

Kaufmann D, Bellver A (2005) Transparenting Transparency: Initial Empirics and Policy Applications. MPRA Paper 8188, University Library of Munich, Germany

Kischel U (2013) Party, pope, and politics? The election of German Constitutional Court Justices in comparative perspective. International Journal of Constitutional Law 11:962–980. https://doi.org/10.1093/icon/mot040

Kleinberg J et al. (2018) Human Decisions and Machine Predictions. The Quarterly Journal of Economics 133(1):273–293. https://doi.org/10.1093/qje/qjx032

Kleinberg J et al. (2019) Discrimination in the Age of Algorithms. Journal of Legal Analysis 10:113–174

Lai C et al. (2014) Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions. Journal of Experimental Psychology General 143(4):1765–1785

Larson et al. (2016) How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 26 June 2020

Levin S (2016) A Beauty Contest Was Judged by AI and the Robots Didn't Like Dark Skin. The Guardian. www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. Accessed 26 June 2020

Licht JD (2011) Do We Really Want to Know? The Potentially Negative Effect of Transparency in Decision Making on Perceived Legitimacy. Scandinavian Political Studies 34:183–201. https://doi.org/10.1111/j.1467-9477.2011.00268.x

Licht JD (2014) Policy area as a potential moderator of transparency effects: An experiment. Public Administration Review 74(3):361–371

Licht KD, Licht JD (2020) Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. AI & Society 35(4):917–926. https://doi.org/10.1007/s00146-020-00960-w

Liptak A (2008) U.S. voting for judges perplexes other nations. The New York Times. www.nytimes.com/2008/05/25/world/americas/25iht-judge.4.13194819.html. Accessed 26 June 2020

Lockhart T (2000) Moral uncertainty and its consequences. Oxford University Press, Oxford

Lowry S, Macpherson G (1988) A blot on the profession. British Medical Journal 296(6623):657–658. https://doi.org/10.1136/bmj.296.6623.657

Macaskill W (2014) Normative Uncertainty. Dissertation, Oxford University

MacAskill W, Bykvist K, Ord T (2020) Moral Uncertainty. Oxford University Press, Oxford

Martinez E, Winter CK (2021) Artificial Intelligence in the Judiciary: A Survey of Expert Opinion. [Manuscript in preparation]

McGinley A (2011) Ricci v. DeStefano: Diluting Disparate Impact and Redefining Disparate Treatment. Nevada Law Journal 12(3):626–639

Meadow W, Sunstein C (2001) Statistics, Not Experts. Duke Law Journal 51:629–646

Michaels AC (2019) Artificial Intelligence, Legal Change, and Separation of Powers. University of Cincinnati Law Review 88:1083–1103

Moore S (2018) Towards a Sociology of Institutional Transparency: Openness, Deception and the Problem of Public Trust. Sociology 52(2):416–430. https://doi.org/10.1177/0038038516686530

Müller VC, Bostrom N (2016) Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller VC (ed) Fundamental Issues of Artificial Intelligence. Springer International Publishing, Cham, pp 555–572

Mukand S, Rodrik D, (2020) The Political Economy of Liberal Democracy. The Economic Journal 130(627):765–792. https://doi.org/10.1093/ej/ueaa004

Niiler E (2019) Can AI Be a Fair Judge in Court? Estonia Thinks So. Wired. www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so. Accessed 25 June 2020

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Broadway Books, New York

Oswald F et al. (2013) Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. Journal of Personality and Social Psychology 105(2):171–192. https://doi.org/10.1037/a0032734

Pejovic C (2001) Civil Law and Common Law: Two Different Paths Leading to the Same Goal. Victoria University of Wellington Law Review 32(3):817–842

Piccinini G (2016) The Computational Theory of Cognition. In: Müller V. (ed) Fundamental Issues of Artificial Intelligence. Synthese Library, vol 376. Springer, Cham, pp 203–221

Pinker S (2011) The Better Angels of Our Nature: Why Violence Has Declined. Viking Press, New York

Pinker S (2018) Enlightenment Now: The Case for Reason, Science, Humanism, and Progress. Viking Press, New York

Piovesan C, Ntiri V (2018) Adjudication by algorithm: The risks and benefits of artificial intelligence in judicial decision-making. The Advocates' Journal 44:42–45

Prunkl C, Whittlestone J (2020) Beyond Near- and Long-Term: Towards a Clearer Account ofResearch Priorities in AI Ethics and Society. arXiv:2001.04335v2 [cs.CY]

Przeworski A (2000) Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990. Cambridge University Press, Cambridge

Rachlinski J, Wistrich A (2017) Judging the Judiciary by the Numbers: Empirical Research on Judges. Annual Review of Law and Social Science 13:203–229. https://doi.org/10.1146/annurev-lawsocsci-110615-085032

Reiling D (2018) What role for AI in a judge's decision-making process? Lecture presented at the European Commission for the Efficiency of Justice's Conference on "Artificial Intelligence at the Service of the Judiciary." Council of Europe. www.coe.int/en/web/cepej/justice-of-the-future-predictive-justice-and-artificial-intelligence. Accessed 25 June 2020

Richard M, Solow-Niederman (2019) Developing Artificially Intelligent Justice. Stanford Technology Law Review, 22(2):242–289

Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. Journal of Risk and Uncertainty 1:7–59

Schumpeter (1942) Capitalism, Socialism and Democracy. Harper & Brothers, New York

Shaked-Schroer N (2008) Reducing racial bias in the penalty phase of capital trials. Behavioral Sciences and the Law 26(5):603–617

Sidgwick H (1907), The Methods of Ethics, 8th Edition

Smith M (2016) In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures. The New York Times. www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html. Accessed 26 June 2020

Sourdin T (2018) Judge v Robot? Artificial Intelligence and Judicial Decision-Making. University of New South Wales Law Journal 41(4):1114–1133

Sourdin T, Cornes R (2018) Do Judges Need to Be Human? The Implications of Technology for Responsive Judging. In: Sourdin T, Zariski A (eds) The Responsive Judge: International Perspectives. Springer, Singapore, pp 87–119

Struchiner N, Almeida G, Hannikainen I (forthcoming 2020) Legal Decision-Making and the Abstract/Concrete Paradox. Cognition

Sunstein (2019) Algorithms, Correcting Biases. Social Research: An International Quarterly 86(2):499–511

Tarsney C (2018) Moral Uncertainty for Deontologists. Ethic Theory Moral Prac 21:505–520. https://doi.org/10.1007/s10677-018-9924-4

Teal C et al. (2012) Helping medical learners recognise and manage unconscious bias toward certain patient groups. Medical Education 46(1):80–88

Tversky A, Kahneman D (1973) Availability: A heuristic for judging frequency and probability. Cognitive Psychology 5:207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Volokh E (2019) Chief Justice Robots. Duke Law Journal, 68:1135–1192

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 7(2):76–99

Wallace M et al. (2017) Separate and unequal: Structural racism and infant mortality in the US. Health & Place 45(3):140–144

Waltl B, Vogl R (2018) Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. Datenschutz und Datensicherheit 42:613–617

Wang P, Goertzel B (2007) Introduction: Aspects of Artificial General Intelligence. In: Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms, Proceedings of the AGI Workshop 2006

Winter CK (2021) Metamoralisches Strafrecht. [Manuscript in preparation]

Winter CK (2020) The Value of Behavioral Economics for EU Judicial Decision-Making. German Law Journal 21(2):240–264

Winter CK, Schuett J, Martinez E, Van Arsdale S, Araújo R, Hollman N, Sebo J, Stawasz S, O'Keefe C, Rotola G (2021) Legal priorities research: A research agenda. Legal Priorities Project. https://www.legalpriorities.org/research_agenda.pdf. Accessed 10 January 2021

Wistrich A et al. (2015) Heart Versus Head: Do Judges Follow the Law or Follow Their Feelings? Texas Law Review 93:855–923

Wistrich A, Rachlinski J (2018) Implicit Bias in Judicial Decision Making, How It Affects Judgment and What Judges Can Do About It. In: Redfield S (ed) Enhancing Justice, Reducing Bias. ABA Book Publishing, Chicago, pp 87–130

World Government Summit (2018) Could an AI ever replace a judge in court? www.worldgovernmentsummit.org/observer/articles/could-an-ai-ever-replace-a-judge-in-court. Accessed 25 June 2020

Završnik A (2020) Criminal justice, artificial intelligence systems, and human rights. ERA Forum 20:567–583. https://doi.org/10.1007/s12027-020-00602-0

Zavyalova V (2018) Save money on legal advice: AI is replacing lawyers in Russia. Russia Beyond. www.rbth.com/science-and-tech/327585-free-legal-advice-robolawyer. Accessed 25 June 2020

Zhang M (2015) Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software. Forbes. www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/#61fdae0c713d. Accessed 26 June 2020