scripted

Volume 18, Issue 1, September 2021

Legal Algorithms and Solutionism: Reflections on Two Recidivism Scores

*Marc Mölders**

DOI: 10.2966/scrip.180121.57

Abstract
Algorithms have entered courts, e.g. via scores assessing recidivism. At first sight, recent applications appear to be clear cases of solutionism, i.e. attempts at fixing social problems with technological solutions. Deploying thematic analysis on assessments of two of the most prominent and widespread examples of recidivism scores, COMPAS and the PSA, casts doubt on this notion. Crucial problems – as different as "fairness" (COMPAS) and "proper application" (PSA) – are not tackled in a technological manner but rather by installing conversations. It shows that even technorationalists never see the technological solution in isolation but are actively searching for flanking social methods thereby accounting for problems that cannot be eased technologically. Furthermore, we witness social scientists called upon as active parts of such engineering.

Keywords
algorithms, fairness, pretrial, recidivism, risk assessment, solutionism

* Senior Lecturer, Law & Society Unit, Faculty of Sociology, Bielefeld University, Germany, marc.moelders@uni-bielefeld.de.

1 Introduction

Algorithms have entered courts. Scores and legal algorithms assessing recidivism or failure to appear in court are both widespread and widely criticised. Some confusion might be due to rather vague notions of what algorithms are.[1] To clarify this term from the very beginning, the contribution at hand understands algorithms as step-by-step instructions to solve a defined (respectively: definable) problem in finite time by a computer, written in computer language, i.e. code.[2]

The algorithms this article deals with do not replace human decision-making but assist it. The difference the algorithm makes does not necessarily need to result in deviant decisions (such as detain or release). The main difference rather refers to the way to get to decisions – especially with respect to the pace and the number of information to process.

Solutionism is a critical stance putting forward that we live in an era which is getting used to technical solutions – especially to algorithms – for social problems. After discussing "sentencing information systems" as legal algorithm's predecessors, the concept of solutionism is sketched in section 2. Algorithms quantifying a defendant's likelihood to reoffend seem to be clear cases to criticise as solutionism. Launching such technological solutions, this is at the core of my argument, produces problems for which not even "technorationalists" demand technical, but rather social strategies. Unlike technoescapists, technorationalists do see the need for politics and the law as social systems, yet they consider such systems underperforming. According to

---

1 Robert Seyfert and Jonathan Roberge (eds.), *Algorithmic cultures. Essays on meaning, performance and new technologies* (London, New York: Routledge, 2006).

2 Angèle Christin, "Algorithms in practice. Comparing web journalism and criminal justice" (2017) 4(2) *Big Data & Society* 1-14, p. 2.

them, such obstacles might best be removed by technology. But even this certain type of solutionist never sees the technological solution in isolation, but is surprisingly willing to advocate for flanking social and organisational methods to get the techno-solution work properly, or argue for controlled remedial action on the human side when the project runs into trouble.

The article arrives at this main conclusion by deploying thematic analysis which is introduced in section 3.1. Two types of material – self-descriptions from the producers and evaluative studies – are analysed this way. Sections 3.2 and 3.3 present the results referring to two of the most prominent examples for legal algorithms or recidivism scores: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) and the Public Safety Assessment (PSA). Problems becoming increasingly visible after their launch refer to their fairness across different societal groups as well as their (proper) usage across different parts of the courtroom workgroup. As different as these problems might seem, it is striking that corresponding solutions, not least proposed by technologists, call for social, conversational, non-technical solutions. Thus, this study adds to the literature on artificial intelligence (AI) in the justice system by deploying thematic analysis to arrive at a more nuanced understanding of solutionism in the justice system, and the explanatory value of this concept.

The observations from the two examples are discussed in section 4. Working on the human rather than technical capacity to be open to surprise is done in a controlled, yet non-technical manner. Fairness, it is called upon, should be discussed in formats of stakeholder participation. On the one hand, this demand is supported by a wide range of experts. On the other hand, there seems to be little interest in society on algorithms and their regulation beyond expert elites; sparking public interest in this topic seems hard to engineer, again.

Section 5 summarises, draws some final conclusions, and reflects on a rather surprising and interesting demand for social scientific expertise for

working on problems technology poses when it is built to work on social problems.

2 Solutionism: Conceptual context and thesis

Section 3 will introduce two of the most prominent examples for legal algorithms or recidivism scores: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) and the Public Safety Assessment (PSA). Here, assessing recidivism means that a trained algorithm scans data from criminal records or from interviews by probation officers to develop a score intended to assist judicial decisions. Differences between both options will be discussed later.

These are comparatively recent developments in terms of technology. Yet this does not mean it is coming without precedent. It is worth a reminder of the discussion around and the practice of sentencing grids. In his evaluation of a federal sentencing grid, Miller (1991) also analysed numbers' capacity to aid judicial decision-making in the United States.[3] This federal grid showed forty-three rows (Offense Levels) and six columns (Criminal History Levels) producing 258 "boxes" altogether. Although we are talking about sheets of paper, it is easy to see this as an attempt to "automate" sentencing. Miller's conclusion was not to compute this process, but to simplify it. The Minnesota grid being the role model of that time, the author comes up with seven offense levels.[4] We will see that it is particularly remarkable that this experimental grid did not include a criminal history dimension.

---

[3] Marc Miller, "True Grid: Revealing Sentencing Policy" (1991) 25(3) *U.C. Davis Law Review* 587-615.

[4] See Andrew von Hirsch, Kay A. Knapp, and Michael H. Tonry, *The Sentencing Commission and Its Guidelines* (Boston: Northeastern University Press, 1987). For an update on federal sentencing schemes see Dawinder S. Sidhu, "Towards the Second Founding of Federal Sentencing" (2017) 77(2) *Maryland Law Review* 485-546.

This discussion shows that there is a long tradition of trying to assist judicial sentencing with numbers, cutting discretion in favour of quantified standardisation.[5] It may be due to such practical interventions into the profession that almost all such projects were accompanied by judicial resistance. Tata et al. (1996) describe different Sentencing Information Systems (SIS) worldwide.[6] In Canada, for instance, judges showed little interest in information about current court practice. They were not accustomed to using numerical information. In another Canadian case, in British Columbia, it was concluded that judges were insufficiently consulted and involved, particularly in the project's early stages.

Exactly this was changed in a New South Wales SIS. Here, judicial education and training was part of the projects from the very beginning. Furthermore, this was organised by the same commission that managed the SIS. This was reported to have been well-received by users and to have served as a role model for the Scottish SIS.[7]

Thus, we know of a long history of numerical assistance as well as of judicial resistance against it. This leads to the question, what – if anything – is new about algorithmic support systems. To state the obvious, such software can proceed more data faster and spot formerly overseen patterns. Maybe it is more important to emphasise that now there is technology to match the well-

---

5 See Leslie T. Wilkins et al., *Structuring Guidelines: Structuring Judicial Discretion. Report on the Feasibility Study* (U.S. Justice Department, 1978).

6 Cyrus Tata, John N. Wilson, and Neil Hutton, "Representations of Knowledge and Discretionary Decision-Making by Decision-Support Systems: The Case of Judicial Sentencing" (1996) (2) *The Journal of Information, Law and Technology*.

7 Janet Chan, "A Computerised Sentencing System for New South Wales Courts" (1991) (137) *Computer Law and Practice*; Neil Hutton, Cyrus Tata, and John N. Wilson, "Sentencing and Information Technology: Incidental Reform?" (1995) 2(3) *International Journal of Law and Information Technology*. For an overview of the contributions of digital technologies, AI to both the legal professions and the police, see Ephraim Nissan, "Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement" (2017) 32(3) *AI & Society* 441-464.

documented will to simplify judicial decision-making. Moreover, some stances would hold that this refers to a general trend to search for technological aid for all sorts of problems: solutionism.

This article's empirical part will mainly deal with recidivism scores. Whatever the problem might be that such algorithms are meant to solve – be it overcrowded jails and prisons, the unjust bail system or the limited human capacity of information processing as well as human biases – they all fit to the well-received diagnosis of "solutionism". Solutionism is a critical concept which Evgeny Morozov defines as follows: "Recasting all complex social situations either as neatly defined problems with definite, computable solutions or as transparent and self-evident processes that can be easily optimized – if only the right algorithms are in place! […] I call the ideology that legitimizes and sanctions such aspirations 'solutionism'".[8]

Pretrial hearings dedicated to deciding whether a defendant is released or detained might be viewed as complex and delicate trade-offs between personal freedom and social security. For an algorithm, there is no such thing as a delicate trade-off as long as there is a running script and sufficient data. In Morozov's terms, recidivism scores would rather appear as projects by technorationalists than by technoescapists. While technoescapists plan to get rid of established (legal and political) institutions altogether by technology, technorationalists intend to use technology to repair the current system: "technorationalists do not aim to rid us of building codes-they, like good technocrats, would prefer that such codes were adopted swiftly, without too much unnecessary consultation and debate".[9]

---

[8] Evgeny Morozov, *To Save Everything, Click Here. Technology, Solutionism and the Urge to Fix Problems that Don't Exist* (New York: PublicAffairs, 2013), p. 5.

[9] *Ibid.*, p. 132.

We cannot prove what technorationalists actually prefer. What the paper at hand intends to assess instead is that the introduction of computable solutions – such as the legal algorithms under scrutiny here – is accompanied and followed by much "consultation and debate". Solutionism has gained traction in social theory.[10] Numbers showing the likelihood of reoffending instead of time-consuming interviews might be considered as a showpiece for a solutionist project. The argument at hand will not disagree but claims that the solutionism critique stops (too) early. This critique's focus on algorithms as solvers of social problems leads a) to overlook problems occurring when algorithms are in place and b) to conceive of technorationalists as naïve and solely believing in technology's capacity for rationalisation.

Thus, the thesis is formulated that crucial problems – as different as "fairness" (COMPAS) and "proper application" (PSA) – are not tackled by technological but rather by social means. In the cases introduced in the following section, we might even speak of installed conversations.

3 Life after launch: Two cases of recidivism scores

3.1 Methodological remarks

Methodologically, this examination relies on thematic analysis[11] of two kinds of material: a) self-descriptions stemming from the producers and b) evaluative studies, largely coming from academia with the significant exception of one

[10] For more references see Oliver Nachtwey and Timo Seidl, "The Solutionist Ethic and the Spirit of Digital Capitalism" (2020), available at https://doi.org/10.31235/osf.io/sgjzq (accessed 24 July 2020).

[11] Jennifer Fereday and Eimear Muir-Cochrane, "Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development" (2006) 5(1) *International Journal of Qualitative Methods* 80-92.

report from investigative journalists.[12] Material A may come from websites, newspapers, magazines, newsletters, or podcasts – all of which are, in principle, publicly available. Whenever a conclusion is drawn from such a site, it is cited. Material B entails evaluations, assessments, and research papers.[13] This research design offers the advantage to draw conclusions from visiting both sides, proponents' internal view as well as (critical) appraisals from outside.

Asking for what problems occur after legal algorithms have been launched and what corresponding solutions are discussed, calls for a theory-driven research design that leaves space for insights offered by the data. Therefore, thematic analysis seems appropriate as it seeks a balance of deductive coding, derived from the theoretical framework and inductive coding from themes emerging from the material. Themes are identified through careful (re-)reading of the data as a form of pattern recognition, where emerging themes become the categories for analysis.[14] According to Boyatzis, a theme is "a pattern in the information that at minimum describes and organises the possible observations and at maximum interprets aspects of the phenomenon".[15] This approach is complemented by deductively scanning the data for answers to questions relevant from a certain conceptual perspective, such as: What problems occur after legal algorithms have been launched and what corresponding solutions are discussed? That fairness and (appropriate) usage, as we will see in the following, are the essential problems dealt with is a result of this kind of

---

12 Julia Angwin et al., "Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks" (2016), available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 24 July 2020).

13 See Appendix for this collection.

14 Fereday and Muir-Cochrane, *supra* n. 11, p. 82.

15 Richard E. Boyatzis, *Transforming qualitative information. Thematic analysis and code development* (Thousand Oaks, London: SAGE), p. 161.

approaching the data. Coming to such a result then leads to the next phase in which the material can be scanned for contributions to these more specific problems and solutions.

3.2 Case A: COMPAS

COMPAS stands for "Correctional Offender Management Profiling for Alternative Sanctions". This name hints at the fact that it was not planned to be used in courts initially.[16] Yet it is used as a risk assessment instrument to predict, *inter alia*, recidivism in pretrial hearings frequently dealing with bail decisions. The private company Northpointe (now: Equivant) developed this algorithm from answers to 137 questions stemming from interviews by probation officers or from criminal records.[17] However, Equivant states that their pretrial risk assessment algorithm uses only six features.[18] Protected categories such as "race" are not surveyed explicitly but there are a lot of questions targeting the defendant's environment, so-called extra-legal factors. The software calculates a score between one and ten which is thought to assist judicial decision-making.

COMPAS became more widely known through the publication of Julia Angwin and her colleagues from the investigative journalism newsroom ProPublica: "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks" in 2016. This study basically dealt with the statistical evidence that COMPAS has a racial bias.[19] It says:

---

16 See https://www.equivant.com/compas-classification/ (accessed 24 July 2020).

17 Julia Dressel and Hany Farid, "The accuracy, fairness, and limits of predicting recidivism" (2018) 4(1) *Science advances* 1-5.

18 Sam Corbett-Davies and Sharad Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning" (2018), available at https://arxiv.org/pdf/1808.00023 (accessed 24 July 2020), p. 19.

19 Angwin et al., *supra* n. 12.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labelling them this way at almost twice the rate as white defendants: 44.9: 23.5%.
- White defendants were mislabelled as low risk more often than black defendants: 28: 47.7%.

Thus, the rationale is an unjust distribution of false positives and false negatives. Interestingly, other studies that used the same data came to opposite conclusions; precisely that COMPAS was a fair instrument precisely because of a lack of difference in predictive utility by race.[20] Here, we witness two mathematically incompatible notions of fairness.[21] This made Science & Technology Studies (STS) scholars Peter Müller and Nikolaus Pöchhacker conclude that all related discussions were located on a statistical level. It was no longer about whether using such instruments was the right way but only *how* to quantify risk assessment. Supporting as well as criticising legal algorithms was done in terms of mathematics, leaving anything incalculable aside.[22] This would mean that technology also dominated an algorithm's life after launch.

As shown, fairness and equal treatment were the aspects that lead to especially controversial discussions around COMPAS. Richard Berk and his

---

[20] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks" (2016) 80(2) *Federal Probation* 38-46.

[21] Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments" (2017) 5(2) *Big Data* 153-163; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores" (2016), available at https://arxiv.org/pdf/1609.05807 (accessed 28 July 2020); Jon Kleinberg et al., "Discrimination in the Age of Algorithms" (2019), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669 (accessed 28 July 2020).

[22] Peter Müller and Nikolaus Pöchhacker, "Algorithmic Risk Assessment als Medium des Rechts. Medientechnische Entwicklung und institutionelle Verschiebungen aus Sicht einer Techniksoziologie des Rechts" *Österreichische Zeitschrift für Soziologie* 44(S1) 157-179, p. 168.

colleagues from Statistics, Criminology, and Computer and Information Science have shown that there are even five different definitions of algorithmic fairness: Overall accuracy equality, statistical parity, conditional procedure accuracy equality, conditional use accuracy equality, and treatment equality.[23] Although the authors develop an overarching concept called "total fairness", they insist that this was practically unattainable. Generally speaking, accuracy and fairness are conflicting goals.

Berk et al. conclude that "in the end, it will fall to stakeholders – not criminologists, not statisticians and not computer scientists – to determine the tradeoffs. […] These are matters of values and law, and ultimately, the political process. They are not matters of science".[24] To work on the problem of fairness, they propose procedures of stakeholder participation. Obviously, this is not a technical solution. At most, we could describe this as social engineering which may be followed by technology: "If there is a policy preference, it should be built into the algorithm".[25] It seems to be far from clear, what deliberation format would be the first choice: "In some cases, this can be through deliberations of a city council or other legislative bodies. In other cases, reform efforts are guided by a standing committee of stakeholder representatives, a commission, or an official advisory board. In yet other cases, there can be an ad hoc oversight committee with wide stakeholder representation. In practice, there are a host of details to be worked out such as which stakeholders can participate and the

---

23 Richard Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art" (2018) 50(1) *Sociological Methods & Research* 3-44.

24 *Ibid.*, p. 35.

25 *Ibid.*, p. 31.

procedural rules to be adopted. Further discussion would be a lengthy diversion and beyond the expertise of the authors".[26]

Müller and Pöchhacker had concluded that the entire fairness debate was using statistics as a kind of official language.[27] Here, we witness that native speakers of exactly this language rather point to languages that solutionists must qualify as outdated: the languages of the law and politics.

If we agreed that this casts doubt on the notion that it is all about technology, we can now turn to the second case: The PSA. As will be shown, we can conceive of the PSA as a case of learning from the experiences with COMPAS in several ways.

3.3 Case B: PSA

PSA stands for Public Safety Assessment. Although this was not planned from the very beginning, all factors this algorithm uses can be seen on a website.[28] Unlike Northpointe, philanthropists Laura and John Arnold – respectively their organisation Arnold Ventures (formerly: Laura and John Arnold Foundation) – did not create a for-profit product.[29] These differences can be regarded as first attempts at working on popular accusations like a lack of transparency and profit-orientation. To tackle COMPAS' main problem – fairness – social science was commissioned; Arnold Venture is describing itself as an evidence-based endeavour. As its guiding principle, the organisation proclaims: "to invest in

---

[26] Richard Berk and Ayya A. Elzarka, "Almost politically acceptable criminal justice risk assessment" (2020) *Criminology & Public Policy* 1-27.

[27] Müller and Pöchhacker, *supra* n. 22, p. 168.

[28] See https://advancingpretrial.org/psa/factors/ (accessed 24 July 2020).

[29] See Marc Faddoul, Henriette Ruhrmann, and Joyce Lee, "A Risk Assessment of a Pretrial Risk Assessment Tool: Tussles, Mitigation Strategies, and Inherent Limits" (2020), available at http://arxiv.org/pdf/2005.07299v1 (accessed 30 July 2021).

evidence-based solutions that maximize opportunity and minimize injustice."[30] Social scientists[31] were asked to examine what factors – compatible with principles of equality – are most predictive of new (violent) criminal activity and failure to appear. A data set of 1.5 million cases of which approximately 750,000 cases were analysed from roughly 300 jurisdictions across the United States lead to these nine factors: Current violent offense, pending charge at time of the offense, prior misdemeanour conviction, prior felony conviction, prior violent conviction, prior failure to appear pretrial in past two years, prior failure to appear pretrial older than two years, prior sentence to incarceration, and age at current arrest; this last one being the only so-called extra-legal factor. Each factor adds a specific value to the overall risk score, which is then scaled down to separate scales for "Failure to Appear" (FTA) and "New Criminal Activity" (NCA) that range from one to six. Furthermore, there is a "New Violent Criminal Activity" (NVCA) flag (i.e., binary indicator of yes/no).[32]

Arguably, it is more interesting what the PSA does not look at: Race, gender, income, education, home address, drug use history, family status, marital status, national origin, employment, and religion.[33] By now, it is used in the states of Arizona, Kentucky, New Jersey, and Utah, as well as in counties and large cities, such as Phoenix, Chicago, and Houston.[34] Being aware of the problems COMPAS was confronted with (respectively: the software confronted

---

30 See https://www.arnoldventures.org/about (accessed 24 July 2020).

31 Marie Van Nostrand has earned a Master's Degree in Public Administration, a second Master's degree in Urban Studies and a Doctorate in Public Policy with a specialty in research methods and statistics; Christopher Lowenkamp is a Social Science Analyst.

32 Matthew DeMichele et al., "The Public Safety Assessment. A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky" (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168452 (accessed 28 July 2020), p. 18.

33 See http://nacmconference.org/wp-content/uploads/2014/01/A2JLab-BackgroundMaterials-20170130.pdf (accessed 24 July 2020).

34 See https://advancingpretrial.org/psa/psa-sites/ (accessed 24 July 2020).

others with), AV came up with a detailed PSA implementation process which is divided into seven phases: Readiness, Engagement, Assistance, Assessment, Automation, Training, Fidelity. These seven phases contain sixteen different steps altogether. [35]

The details this implementation process displays already hint at the fact that AV and its executing grantees – Advancing Pretrial Policy and Research (APPR) – know about implementation or application problems. This means that it is not relied on the properties of the algorithm but that a need for adapting to possible users is recognised. Again, this knowledge comes from social science. In her ethnographic study on the actual use of the PSA in several jurisdictions, Angèle Christin found instructive buffering strategies judges developed in order to lower the impact of this technology on their routines. For example, some printed risk score sheets to place them towards the end of the hundred pages or more that made up the files.[36]

While we cannot know whether AV is aware of this certain study, the philanthropists did work with social and data scientists to constantly evaluate both, the actual use as well as the views of their instrument by courtrooms

[35] See https://advancingpretrial.org/implementation/overview/ (accessed 28 July 2020).

[36] Christin, *supra* n. 2, p. 9.

professionals.[37] Here, it is explicitly referred to organisational sociology.[38] Thus, AV and its partners do not only strive for the best technology but weigh in such things as "courtroom culture". A "courtroom workgroup", this is taken directly from organizational sociology, was not made up of judges, prosecutors, and defenders, but must include pretrial officers which usually were overlooked: "pretrial staff have not traditionally been classified a member of the courtroom work group but their role of completing pretrial risk assessments for defendants can contribute to other legal actors' holistic understanding about a defendant and potentially shape collective discretion within the courtroom work group. Without a shared understanding of the utility of the tool, its value and use may be compromised".[39] Such a shared understanding in terms of informal and implicit rules is what the term "organizational culture" refers to.

Because this evaluation study is listed on the operating organization's website, it is likely that it was considered. Conversely, a reliance on the benefits of machine learning alone becomes rather unlikely. Such a stance is proposed by Thomas et al., namely to search for machine learning algorithms "that provide

---

[37] Matthew DeMichele et al., "What Do Criminal Justice Professionals Think About Risk Assessment at Pretrial?" (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168490 (accessed 28 July 2020); Matthew DeMichele et al., "The Intuitive-Override Model. Nudging Judges Toward Pretrial Risk Assessment Instruments" (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168500 (accessed 28 July 2020); Matthew DeMichele et al., "The Public Safety Assessment. A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky" (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168452 (accessed 28 July 2020); Megan T. Stevenson, "Assessing Risk Assessment in Action" (2018) (103) *Minnesota Law Review* 303-384.

[38] Paul J. DiMaggio and Walter W. Powell, "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields" (1983) 48(2) *American Sociological Review* 147-160.

[39] Matthew DeMichele et al., "What Do Criminal Justice Professionals Think About Risk Assessment at Pretrial?" (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168490 (accessed 28 July 2020), p. 4.

their users with the ability to easily (that is, without requiring additional data analysis) place limits on the probability that the algorithm will produce any specified undesirable behavior".[40] Will Knight holds that this approach is hardly able to solve the problem of algorithms misbehaving, not least "because there's no guarantee organizations deploying AI will adopt such approaches when they can come at the cost of optimal performance".[41] This hints at the problem that any novelty has to be translated within organizations, another considerable aspect of an algorithm's life after launch.[42]

Of course, it is not that fairness as a problem has disappeared. Yet, it is striking that the proper use of this risk assessment instrument by courtroom professionals has become central. For every new jurisdiction that declares itself ready to implement the PSA, a Technical Assistance (TA) team is offered helping to adapt the algorithm with regard to both, "standard operating practices and courtroom culture".[43] Thereby, two things are enabled at the same time: On the local level, the algorithm is tailored to the specific needs on-site. Furthermore, this allows the entire project to "scale up." Philanthropists are often confronted with the expectation to make the *world* a better place but then have to work on specific, locally limited projects.[44] Being prepared for adaptations due to

---

40 Philip S. Thomas et al., "Preventing undesirable behavior of intelligent machines" (2019) 366(6468) *Science* 999-1004, p. 1003.

41 Will Knight, "Researchers Want Guardrails to Help Prevent Bias" (2019), available at https://www.wired.com/story/researchers-guardrails-prevent-bias-ai/ (accessed 24 July 2020).

42 See Barbara Czarniawska and Bernward Joerges, "Travels of Ideas" in Barbara Czarniawska and Guje Sevón (eds.), *Translating organizational change* (Berlin, New York: de Gruyter, 1996), pp. 13-48; Tammar B. Zilber, "Institutional maintenance as narrative acts" in Thomas B. Lawrence (ed.), *Institutional Work: Actors and Agency in Institutional Studies of Organizations* (Cambridge: Cambridge University Press, 2009), pp. 205-235.

43 De Michele et al., *supra* n. 31.

44 This is the core idea of Philanthrocapitalism: Matthew Bishop and Michael Green, *Philanthrocapitalism. How Giving Can Save the World* (New York: Bloomsbury Press, 2008). For problems of organising philanthrocapitalism see Marc Mölders, "Changing the World by

operative as well as cultural differences appears as a "stairway to scalability". Furthermore, this illustrates that scale does not mean size. Scaling up does not equal replicating a standardized solution, but rather to recognize regional and other differences.[45] Again, this does not match the naivety solutionism assumes.

4 Discussion

Both examples discussed do not show machines replacing human (read: judicial) decision-making, we are not (yet) talking about "robot judges".[46] What we do witness, though, is a kind of enhancing human decisions with a certain quality. Of course, to make judges actually use the algorithm (and not use buffering strategies) is risky because it implies communicating that there is something wrong with the way judges routinely act. In section 2 we already saw judge's more general resistance against numerical assistance. What the PSA intends to do is make the (especially seasoned) judge reflect his or her intuition as this is seen as the point at which bias enters. The algorithm should disrupt or irritate this "autopilot mode" and make them reflect their reasoning.[47] An evaluation study called "Nudging Judges Toward Pretrial Risk Assessment Instruments" does not rely on technology design in the first place.[48] Instead, this issue is worked on in meetings and training in which legal professionals are rather talked

---

Changing Forms? How Philanthrocapitalist Organizations Tackle Grand Challenges" (2020), available at https://osf.io/preprints/socarxiv/xh46a/ (Accessed 24 July 2020).

45 Pratima Bansal, Anna Kim, and Michael O. Wood, "Hidden in Plain Sight: The Importance of Scale in Organizations' Attention to Issues" (2018) 43(2) *Academy of Management Review* 217-241, p. 233.

46 But see https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ (accessed 28 July 2020).

47 Marc Mölders, "Irritation expertise. Recipient design as instrument for strategic reasoning" (2014) 2(1) *European Journal of Futures Research* 32.

48 Matthew DeMichele et al., "The Intuitive-Override Model. Nudging Judges Toward Pretrial Risk Assessment Instruments" (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168500 (accessed 28 July 2020).

than nudged into proper application ("researcher-judge feedback loops"). This is even more remarkable as such face-to-face interactions are precisely what the algorithm is meant to replace: the defendant interview.

By far, this conclusion is not only drawn in a single evaluation study. Non-profit organisation "Partnership on AI" issued a report that recommends „Users of risk assessment tools must attend trainings on the nature and limitations of the tools".[49] Even studies supporting the claim that algorithmic risk assessments can often outperform human predictions of reoffending emphasise non-technical opportunities. Because the typical justice setting hardly offers feedback opportunities, judges usually never find out what happens to those they sentenced. Therefore, Lin et al. put forward that jurisdictions could create a learning environment by requiring that judges express and record their intuitive estimates of risk and by providing regular feedback on past predictions. "With that information, judges could, for example, see the actual postrelease recidivism rate of those that they had deemed 'high risk.' […] [T]his feedback could correct tendencies to overpredict recidivism".[50] Understood this way, feedback is conceived of as a social or organisational rather than a technical solution for the problem of a judges' overreliance on his or her routine.

Thus, it is worked on the human rather than technical irritability. Therefore, exchanges between judges and these algorithms cannot be treated as

---

[49] Partnership on AI, "Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System" (2019), available at https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/ (accessed 24 July 2020), p. 26. See also Jennifer Skeem, Nicholas Scurich, and John Monahan, "Impact of risk assessment on judges' fairness in sentencing relatively poor defendants" (2020) 44(1) *Law and Human Behavior* 51-59.

[50] Zhiyuan "Jerry" Lin et al., "The limits of human predictions of recidivism" (2020) 6(7) *Science advances* 1-8, p. 5.

"artificial communication" according to Elena Esposito (2017).[51] This is given when "the interlocutor (…) has a sufficiently complex structure for the interaction to produce information different from what the user already knows, and this information is attributed to the partner."[52] It is about "the production of appropriate and informative surprises".[53] To come as a surprise, though, there are trainings and meetings – not technology in a narrower sense. Designing trainings and meetings rather resembles social engineering.[54] In a similar vein, Stefan Meißner (2017) defines technology as a scheme used to observe the world according to the distinction of controllable/uncontrollable. This allows him to conceptually incorporate "technologies of the social", e.g. the installation of conversation techniques to attempt at controlling stubborn human beings.[55]

Elaborating on the problem of fairness and possibly corresponding solutions lead to comparable results. Technologists and non-technologists seem to agree that fairness is not just hard to translate into code but is a matter of law and politics in the first place. This is not to say that there is no such thing as proposals for more or less purely technological solutions. Several aspects within the debate around "explainability" could serve as examples. Here, algorithms are to be developed that not only give the reasons for a certain result but come up with recommendations on how to change it ("counterfactual explanations").[56]

---

51 Elena Esposito, "Artificial Communication? The Production of Contingency by Algorithms" (2017) 46(6) *Zeitschrift für Soziologie* 249-265.

52 *Ibid.*, p. 258.

53 *Ibid.*, p. 257.

54 On social engineering see Thomas Etzemüller, *Alva and Gunnar Myrdal. Social Engineering in the Modern World* (Lanham: Lexington Books, 2014).

55 Stefan Meißner, *Techniken des Sozialen. Gestaltung und Organisation des Zusammenarbeitens in Unternehmen* (Wiesbaden: Springer VS, 2017); my translation.

56 Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual Explanations without Opening the Black Box. Automated Decisions and the GDPR" (2018) 31(2) *Harvard Journal of Law & Technology* 31 (2) 841-887. The same authors stress the importance that the technical and the legal community learn from each other: The use and limits of interpretation and the use and limits of statistics. Therefore, they "propose summary statistics that describe 'conditional

"Actionable Recourse" as a model adds to that an exclusive focus on modifiable results in terms of changeable/unchangeable (e.g. college degree vs. gender).[57] We might view a college degree as a changeable factor, but if gender or age show a massive impact on a certain result, things turn out differently. Recently, Sandra Wachter and Brent Mittelstadt took an additional step by discussing a right to "reasonable inferences".[58] Of course, explainability only makes sense when there are reiterative engagements – e.g. credit card applications where one can apply again when rejected. Obviously, this is not the case in sentencing or in decisions concerning recidivism.

Even proposals strongly reminding of technological solutionism are not considered in isolation. Cultural anthropologist Madeleine Clare Elish gets to the heart of it when she puts it this way: "The question of 'What it means for an algorithm to be fair?' does not have a technical answer alone".[59] While this does not come as a surprise from cultural studies, the preceding has shown that even contributions from statistics, computer and information science point to the necessity to look beyond their home ground.

---

demographic disparity' (CDD) […] as a baseline for evidence to ensure a consistent procedure for assessment (but not interpretation) across cases involving potential discrimination caused by automated systems". See Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (2021) 41 *Computer Law & Security Review*, p.6.

57 Alexander Spangher and Berk Ustun, "Actionable Recourse in Linear Classification. Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning" (2018), available at https://econcs.seas.harvard.edu/files/econcs/files/spangher_fatml18.pdf (accessed 28 July 2020).

58 Sandra Wachter and Brent Mittelstadt, "A Right to Reasonable Inferences. Re-Thinking Data Protection Law in the Age of Big Data and AI" (2019) 2019(2) *Columbia Business Law Review* 1-130.

59 This citation is retrieved from https://www.wired.com/story/what-does-a-fair-algorithm-look-like (accessed 28 July 2020). See also, Madeleine Clare Elish and danah boyd, "Situating Methods in the Magic of Big Data and Artificial Intelligence" (2018) 85(1) *Communication Monographs* 57-80.

5 Conclusions

This article set out by casting doubt on the assumption that the spreading use of algorithms within courts serves as a clear case in terms of solutionism. The application of two of the most prominent examples, COMPAS and the PSA, lead to distinct kinds of problems after they have been launched: fairness (COMPAS) and appropriate usage (PSA). Solutions discussed for both were not technological by default. Surprisingly, especially regarding the different nature of these problems, in both cases conversations, such as meetings, training, stakeholder participation, or the like are deemed most appropriate. Even reviving Athenian citizen councils – selecting representatives by lot – is put forward as a model to come to societally sound decisions on algorithms.[60] Apparently, there is a broad consensus towards algorithm regulation. A blind spot seems to be that there is not much interest in algorithms (and their regulation) in both society and politics but that this remains a topic for experts. When even data scientists call for a societal debate, it might be about time to get it started.

Maybe data science's patience is tested too much. It seems as if Richard Berk and his colleagues waited for politics to assume responsibility to involve stakeholders. In the meantime, there are papers proposing "almost politically acceptable" risk assessments.[61] One example for a risk procedure on which a sufficient number of stakeholders agree could (i.e. almost politically acceptable)

---

[60] See Federica Carugati, "A Council of Citizens Should Regulate Algorithms" (2020), available at https://www.wired.com/story/opinion-a-council-of-citizens-should-regulate-algorithms/ (accessed 24 July 2020). The author refers to Federica Carugati, *Creating a constitution. Law, democracy, and growth in ancient Athens* (Princeton: Princeton University Press, 2019); Scott E. Page, *The Difference. How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton: Princeton University Press, 2008).

[61] Berk and Elzarka, *supra* n. 26.

would be to treat all potential offenders as if they are white. Practically this would mean to train the risk algorithm with data from the most privileged group.[62] In another paper, a formal framework is added.[63] In this, the same author remains cognisant about social barriers: "However, the pareto improvement that results must pass political and legal muster before our proposals could properly be implemented. These challenges have yet to be addressed and could well be contentious".[64]

To be precise, comparing technical solutions with social ones, such as deliberative formats, does not mean to waive control usually associated with (working) technology. Instead, we have to expect that training judges in properly using a recidivism score or involving societal stakeholders in procedures is done in a highly controlled manner.

It is this point at which social sciences enter. Discussions around "Public Sociology" demanded from social scientists to get involved when issues of public interest are at stake.[65] The cases discussed rather showed a demand for social scientific expertise from practical sites. Here, sociologists and others were not asked as critical observers but as experts for assessing whether targets are met, for measuring impact, evaluating measures, analysing huge amounts of data, collecting evidence, supporting acceptability, and much more.

---

62 *Ibid*., pp. 20-21.

63 Richard Berk and Arun Kumar Kuchibhotla, "Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Conformal Prediction Sets" (2020), available at https://arxiv.org/abs/2008.11664 (accessed 22 September 2020).

64 *Ibid.*, p. 21.

65 Michael Burawoy, "For Public Sociology" in Dan Clawson et al. (eds.), *Public Sociology: Fifteen Eminent Sociologists Debate Politics and the Profession in the Twenty-first Century* (Berkeley: University of California Press, 2007), pp. 23-66.

Albeit having examined only a tiny fraction, these results call for attention to look twice whether current attempts at making the world a better place actually put technology in the driver's seat.

6 Appendix

6.1 Research Papers on COMPAS

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There is software that is used across the county to predict future criminals. And it is biased against blacks. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533
- Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior*, *36*(1), 21–40. https://doi.org/10.1177/0093854808326545
- Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047
- Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Retrieved from https://arxiv.org/pdf/1808.00023
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), 1-5. https://doi.org/10.1126/sciadv.aao5580
- Fenton, N. E., & Neil, M. (2018). Criminally Incompetent Academic Misinterpretation of Criminal Data - and how the Media Pushed the Fake News. Retrieved from https://www.researchgate.net/publication/322603937
- Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Federal Probation*, *80*(2), 38–46.

- Goel, S., Shroff, R., Skeem, J. L., & Slobogin, C. (2018). The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3306723
- Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*(7), 1-8. https://doi.org/10.1126/sciadv.aaz0652

6.2 Research Papers on PSA[66]

- Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 1-14. https://doi.org/10.1177/2053951717718855
- DeMichele, M., Baumgartner, P., Barrick, K., Comfort, M., Scaggs, S., & Misra, S. (2018). What Do Criminal Justice Professionals Think About Risk Assessment at Pretrial? *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3168490
- DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., & Misra, S. (2018). The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3168452
- DeMichele, M., Comfort, M., Misra, S., Barrick, K., & Baumgartner, P. (2018). The Intuitive-Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3168500
- Redcross, C., & Henderson, B. (2019). Evaluation of Pretrial Justice System Reforms That Use the Public Safety Assessment: Effects in Mecklenburg County, North Carolina. Retrieved from https://www.mdrc.org/publication/evaluation-pretrial-justice-system-reforms-use-public-safety-assessment

[66] For an updated list of research on PSA see https://advancingpretrial.org/psa/research/ (accessed 22 September 2020).

- Stevenson, M. T. (2018). Assessing Risk Assessment in Action. *Minnesota Law Review*. (103), 303–384. https://doi.org/10.2139/ssrn.3016088

6.3 Research Papers on COMPAS and PSA

- Kehl, D. L., & Kessler, S. A. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Retrieved from https://dash.harvard.edu/handle/1/33746041
- Partnership on AI (2019). Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System. Retrieved from https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/