Jailed by a "black box": The impact of opaque algorithms on the right to a fair trial in the United States of America*

Willem Gravett
BLC LLB LLM LLD
Associate Professor, Department of Procedural Law, University of Pretoria

OPSOMMING

Veroordeel deur 'n "black box": Die invloed van ondeursigtige algoritmes op die reg tot 'n billike verhoor in die Verenigde State van Amerika

Ontwikkelings in kunsmatige intelligensie (AI) en masjienleer het daartoe begin lei dat regerings die gesag om openbare funksies te verrig uitkontrakteer aan masjiene. Algoritmiese besluitneming word inderdaad algemeen, van die bepaling van verbruikers se kredietwaardigheid, tot die identifisering van die mees geskikte kandidaat vir 'n betrekking, tot die besluit oor watter matrikulante toegelaat moet word tot 'n universiteit. Afgesien van die breër maatskaplike, etiese, en regsoorwegings, is daar omstredenheid oor die ondeursigtige, onbetwisbare, en ontoerekeningsvatbare aard van AI-stelsels. 'n 2016 beslissing in die Verenigde State van Amerika, *S v Loomis*, illustreer die bedreiging wat die ongereguleerde en onbeperkte uitkontraktering van openbare gesag aan AI-stelsels inhou vir menseregte en die oppergesag van die reg. In hierdie saak het die Hooggeregshof van Wisconsin beslis dat die gebruik van 'n geoutomatiseerde, algoritmiese risikobepaling deur die verhoorhof nie die beskuldigde se reg op 'n billike verhoor geskend het nie, alhoewel die metodologie wat deur die sagteware gebruik is om die risikobeoordeling te maak nòg aan die hof nòg aan die beskuldigde openbaar is. Die betrokke sagteware – 'n sogenaamde getuienis-gebaseerde stelsel – is 'n hoogs omstrede instrument wat verskillende datapunte met betrekking tot 'n beskuldigde in ag neem en dan 'n risikotelling aan sodanige beskuldigde toeken. Hoe hoër die telling, hoe meer waarskynlik word dit geag dat die betrokke beskuldigde in die toekoms weer 'n misdryf sal pleeg. Die algoritme word dus aangewend om die toekomstige gedrag van beskuldigdes te voorspel. Baie beskuldigdes word dan as potensiële toekomstige misdadigers beskou en behandel sonder dat hulle ooit oor die grondslag vir hulle risikotellings ingelig word – en sonder die middele om ooit sodanige grondslag te kan uitvind. Baie van hierdie geoutomatiseerde stelsels, soos COMPAS wat in die *Loomis*-saak ter sprake was, bestaan uit 'n sogenaamde "black box" van privaatsektor-handelsgeheime wat geensins verantwoordbaar is nie.

> "[I]f we're going to use these things and rely on them,
> then let's get as firm a grip [as possible] on how and why
> they're giving us the answers.

---

* This article is based on the author's doctoral thesis *Problematic aspects of the right to bail under South African law: A comparison with Canadian law and proposals for reform* (LLD thesis, University of Pretoria 2000).

> If it can't do better than us at explaining what it's doing,
> then don't trust it."[1]

1 INTRODUCTION

To an ever-increasing degree, Artificial Intelligence (AI)[2] systems and the algorithms that power them are tasked with making crucial decisions that used to be made by humans. Algorithmic decision-making based on big data[3] has become an essential tool and is pervasive in all aspects of our daily lives: the news articles we read, the movies we watch, the people we spend time with, whether we get searched in an airport security line, whether more police officers are deployed in our neighbourhoods, and whether we are eligible for credit, health care, housing, education, and employment opportunities, among a litany of other commercial and government decisions.[4]

Some view this as a cause for celebration. We have come to inhabit a world in which the only sustainable way to make sense of the sheer volume, complexity, and variety of data that are produced daily, is to apply AI.[5] We cede our decision-making to algorithms in large part because of the gains in power, speed, and efficiency that they afford.

However, as algorithms become more accurate predictors, they also become more complex, and, consequently, more opaque and resistant to interrogation.[6] Automated decision-making systems[7] have become "black boxes"[8] – even their designers often do not understand the process through which inputs become outputs. To make matters worse, algorithms often are deliberately shrouded in secrecy as proprietary trade secrets.[9] This opacity prevents those harmed by automated systems from determining either how a decision came about, or the

---

1 Daniel Dennett as quoted in Knight "The dark secret at the heart of AI" *MIT Technology Review* (11 April 2017) available at https://www.technologyreview.com/2017/04/11/5113/-the-dark-secret-at-the-heart-of-ai/ (accessed on 09-06-2020).

2 AI refers to a computer's ability to imitate human intelligent behaviour, especially human cognitive functions, such as the ability to reason, discover meaning, generalise and learn from past experience. Alan Turing defined artificial intelligence as the "science and engineering of making intelligent machines, especially intelligent computer programs." Turing "Computing machinery and intelligence" (1950 *Mind* 433).

3 Big data are extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. These data sets are so large and complex that they are impossible for humans to process, and even difficult or impossible to process using traditional computational methods. See Ishwarappa "A brief introduction on Big Data 5Vs characteristics and hadoop technology" 2015 *Procedia Computer Science* 319–320.

4 Osoba & Welser IV "An intelligence in our image: The risks of bias and errors in artificial intelligence" 2017 *Rand Corporation* 1; Waldman "Power, process, and automated decision-making" 2019 *Fordham LR* 613.

5 Osoba & Welser 6.

6 Waldman 618.

7 Automated decision-making systems use complex mathematical algorithms to identify meaningful relationships and likely patterns in large data sets. Berman "A government of laws and not of machines" 2018 *Boston University LR* 1279.

8 See Pasquale *The black box society: The secret algorithms that control money and information* (2015) 1–17.

9 Wexler "Life, liberty and trade secrets: Intellectual property in the criminal justice system" 2017 *Stanford LR* 11343.

logic or reasoning behind it.[10] According to the legal scholar, Ari Ezra Waldman, the result is[11] –

> "… a technologically driven decision-making process that seems to defy interrogation, analysis and accountability and, therefore, undermines due process. This should make algorithmic decision-making an illegal source of authority in a liberal democracy."

This article examines a previously obscure but rapidly growing area within the field of criminal justice – the use of risk-assessment software, powered by sophisticated and often proprietary algorithms, to predict whether individual criminals are likely to re-offend (the "risk of recidivism"). The focus is on the latest, and perhaps most troubling, use of these risk-assessment tools: their incorporation into the criminal sentencing process.[12] This development raises fundamental legal questions about the right to a fair trial, equal protection, and transparency.

By way of background, several nations have become enthusiastic adopters of automation in the criminal justice system.[13] For example, data analytics and algorithms have for years been applied in the United States criminal justice system in the decision-making processes of law enforcement agencies, corrections officials and judges.[14] The rapid and unprecedented rise of algorithms has been fuelled by a number of factors. In the first instance, vast amounts of data are generated by the ubiquitous use of the internet and smart devices – more data than can be humanly processed – leading to a growing emphasis on data-driven decision-making, not only in our private lives, but also in public policy.[15] Secondly, the algorithmic approach is seen as "cost-effective" in aiding criminal justice officials to prioritise scarce government resources in predicting complex individual behaviours.[16] Thirdly, many believe that these big data-driven algorithmic tools can remove the presence of human adjudicators – and with them their inherent "biases"[17] – from the decision-making process.[18] Justice is dispensed in a more efficient way, so the argument goes, because algorithms rely exclusively on empirical evidence (the "evidence-based" approach), rather than personal judgements.[19]

---

10 Selbst & Barocas "The intuitive appeal of explainable machines" 2018 *Fordham LR* 1091–1101.

11 Waldman 614.

12 See Kehl, Guo & Kessler "Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing" *Responsive Community Initiative, Berkman Klein Center for Internet and Society, Harvard Law School* (2017) 2.

13 Zalnieriute, Bennett Moses & Williams 'The rule of law and automation of government decision-making' 2019 *Modern LR* 1.

14 Liu, Lin & Chen "Beyond *State v Loomis*: Artificial intelligence, government algorithmization and accountability" 2019 *Int'l J L & Information Tech* 124.

15 Kehl, Guo & Kessler 3.

16 Brauneis & Goodman "Algorithmic transparency for the smart city" 2018 *Yale J L & Tech* 114.

17 For a detailed discussion of the biases inherent in algorithmic risk-assessment tools, see, generally, Gravett "Sentenced by an algorithm – Bias and lack of accuracy in risk-assessment software in the United States criminal justice system" 2021 *SA J Crim Justice* (forthcoming).

18 Salman & Le Coz "Race and politics influence judicial debcisions, but Florida's bench is a world of contradictions" *Herald Tribune* (12 December 2016) available at http://projects.-heraldtribune.com/bias/politics/ (accessed on 10-06-2020).

19 Smith "In Wisconsin, a backlash against using data to foretell defendants' futures" *The New York Times* (22 June 2016) available at https://www.nytimes.com/2016/06/23/us/

continued on next page

It is well known that the "tough on crime" policies at both state and federal levels in the United States have mandated long prison sentences for violent and drug-related crimes and repeat offences. The result has been a mass imprisonment problem with unwanted fiscal, social, and political consequences.[20] Escalating corrections costs and the high rate of recidivism have led policymakers to adopt data-driven algorithmic approaches as a move away from heavy reliance on imprisonment.[21]

These innovative technologies claim to inform judges better by profiling offenders based on their risk of recidivism. These technologies take advantage of machine learning[22] algorithms, which generate risk models based on vast quantities of data.[23] Currently, around 60 automated systems have been adopted at every decision point[24] throughout the criminal justice system, from policing to pre-trial bail to determinations of parole to conditions of supervision to post-trial sentencing.[25] However, "the inner workings of these tools are largely hidden from public view".[26]

As yet there does not seem to be any imminent plans to import AI into the South African criminal justice system. On 6 August 2020, the Presidential Commission of the Fourth Industrial Revolution presented its report to the President, and on 6 October the Commission's Report was released for public

---

backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html (accessed on 23-06-2020); Kirchner "Wisconsin court: Warning labels are needed for scores rating defendants' risk of future crime" *ProPublica* (14 June 2016) available at https://www.propublica.org/article/wisconsin-court-warning-labels-needed-scores-rating-risk-future-crime (accessed on 23-06-2020).

20 In the United States, the penal population has quadrupled to a record 2.2 million people currently behind bars from only around 500 000 in the 1980s. This figure represents five times the international average. See Conyers "The incarceration explosion" 2013 *Yale L & Policy Rev* 377–378.

21 Berry-Jester, Casselman a& Goldstein "The new science of sentencing" *Marshall Project* (4 August 2015) available at https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing (accessed on 08-06-2020).

22 Machine learning can be described as the ability of a computer to modify its programming to account for new data and modify its operations accordingly. Machine learning is generally iterative (capable of continually "learning" from new information) and capable of identifying more complex patterns in data. Zalnieriute, Bennett Moses & Williams 9.

23 Kehl, Guo & Kessler 9.

24 Kirchner *ProPublica* (2016).

25 See, for example, Ferguson "Policing predictive policing" 2017 *Washington University LR* 1109–1189; Baradaran & McIntyre "Predicting violence" 2012 *Texas LR* 497–570; Sidhu "Moneyball sentencing" 2015 *Boston College LR* 671–731; Anonymous "The new science of sentencing" *The Marshall Project* (undated) available at https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing (accessed on 18-06-2020). It should be mentioned that these tools were not initially developed for use in sentencing. When risk-assessment tools started being implemented in 1989, they were intended for use in the corrections industry generally – by probation and parole officers – to alert these officials to which individuals to pay the most attention. Chiel "Secret algorithms that predict future criminals get a thumbs up from Wisconsin Supreme Court" *Splinter News* (27 July 2016) available at https://splinternews.com/secret-algorithms-that-predict-future-criminals-get-a-t-1793860613 (accessed on 23-07-2020).

26 Liptak "Sent to prison by a software program's secret algorithm" *The New York Times* (1 May 2017) available at https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html (accessed on 18-06-2020).

consumption.[27] The report speaks broadly of, among other things, investment in human capital, establishing an AI institute, and building 4IR infrastructure. It does not contain any industry-specific recommendations and plans.[28] Also, upon his appointment as Minister of Justice and Constitutional Development, Mr Ronald Lamola emphasised that modernisation of the South African courts would be a priority.[29] However, the focus, at least at present, seems to be on digitising court records and digital case management, rather than incorporating AI technology into the court system.[30] Thus, the scope of this article is limited to the use of risk-assessment software to predict the risk of recidivism of offenders in the United States.

Despite the promise of these systems, the question of how automation interacts with fundamental legal concepts and norms have engendered lively debate among policymakers, politicians, practitioners, and academics in the United States and elsewhere.[31] The ramifications are well illustrated by the decision of the Wisconsin Supreme Court in *S v Loomis*.[32]

2 S v LOOMIS

Loomis is a landmark decision, because it is the first time that a court addressed the legality of using algorithmic risk-assessment tools.[33] Unfortunately, the court only superficially addressed the issue, while leaving aside critical challenges that most courts are likely to face in the age of big data and AI.[34]

---

27 The Commission, consisting of leaders from academia, business and civil society, began its work in May 2019, combining research and stakeholder engagements to generate a comprehensive view of South Africa's current conditions as well as the prospects in the Fourth Industrial Revolution. The Presidency "Presidential commission on 4IR presents recommendations to President Ramaphosa" (6 August 2020) available at http://www.thepresidency.gov.za/press-statements/presidential-commission-4ir-presents-recommendations-president-ramaphosa (accessed 16-11-2020); The Presidency "4IR commission report recommendations gazetted" (62020) available at https://www.sanews.gov.za/south-africa/4ir-commission-report-recommendations-gazetted (accessed on 16-11-2020).

28 Presidential Commission on the Fourth Industrial Revolution *Report, Recommendations and way forward* (March 2020) available at file:///Users/willemgravett/Downloads/4IR-Report-Recommendations-and-Way-Forward.pdf (accessed on 17-11-2020).

29 Davis "Modernising SA courts among Lamola's top priorities" *Eyewitness News* (13 July 2019) available at https://ewn.co.za/2019/07/03/moderinising-sa-courts-among-ronald-lamolas-top-priorities (accessed on 17-11-2020).

30 Ensor "Ronald Lamola commits to modernisation of justice system" *Business Day* (23 July 2020) available at https://www.businesslive.co.za/bd/national/2020-07-23-ronald-lamola-commits-to-modernisation-of-justice-system/ (accessed on 17-11-2020); Makinana "New justice minister Ronald Lamola to digitise SA's paper-laden courts" *Times Live* (3 July 2019) available at https://www.timeslive.co.za/politics/2019-07-03-new-justice-minister-ronald-lamola-to-digitise-sas-paper-laden-courts/ (accessed on 17-11-2020).

31 See, for example, Pasquale (2015). "With increased adoption of these tools," notes the *ABA Journal*, "defense attorneys raise due process concerns, policymakers struggle to provide meaningful oversight, and data scientists grapple with ethical questions regarding fairness and accuracy" Tashea "Risk-assessment algorithms challenged in bail, sentencing and parole decisions" *ABA Journal* (1 March 2017) available at https://www.abajournal.com/magazine/article/algorithm_bail_sentencing_parole (accessed on 18-06-2020). See also Zalnieriute, Bennett Moses & Williams 2.

32 *S v Loomis* 881 NW2d 749 (Wisconsin 2016) 754.

33 Kehl, Guo & Kessler 3.

34 See Liu, Lin & Chen 131.

In 2013, 31-year old Eric Loomis was arrested in La Crosse, Wisconsin, on charges related to a drive-by shooting.[35] Loomis denied any involvement in the shooting, but he nevertheless waived his right to trial and entered a guilty plea to two of the lesser charges – fleeing from a traffic officer and driving a vehicle without the owner's consent.[36] These were all repeat offences. Loomis was also on probation for dealing in prescription drugs, and he was a registered sex offender because of a previous conviction for third degree sexual assault.[37] In mitigation, his attorney emphasised a childhood spent in foster homes where he was abused. With an infant son of his own, Loomis was also training to be a tattoo artist.

Following the plea, the circuit (trial) court ordered a pre-sentencing investigation report, which included a risk-assessment by COMPAS, an algorithmic risk-assessment tool, to aid the court in determining Loomis's sentence. COMPAS assessments estimate the risk of recidivism based on an interview with the accused and information from the accused's criminal history.[38] COMPAS assesses variables under five main areas: criminal involvement, relationships/lifestyles, personality/attitudes, family, and social exclusion.[39]

Because the methodology behind COMPAS is a trade secret, only the software's estimates of recidivism risk are reported to the court.[40] The fact that the COMPAS software is proprietary means that there is no federal oversight, and there is virtually no transparency about its inner workings. As discussed below, COMPAS has created considerable controversy for this very reason.[41]

The COMPAS risk assessment designated Loomis a high risk for all three types of recidivism that the system measured: pre-trial recidivism, general recidivism, and violent recidivism.[42] In imposing the maximum sentence of six years' imprisonment and five years' extended supervision, the judge specifically mentioned the COMPAS score:[43]

> "You are identified through the COMPAS assessment as an individual who is at high risk to the community.... I'm ruling out probation because of the seriousness

35 Loomis was charged with (a) first-degree recklessly endangering safety; (b) attempting to flee or elude a traffic officer; (c) operating a motor vehicle without the owner's consent; (d) possession of a firearm by a felon; and (e) possession of a short-barrelled shotgun or rifle.

36 *Ibid.*

37 *Ibid.*

38 *Ibid.*

39 Specifically, the algorithm uses information from a 137-page survey, separated into several sections, and from the accused's public criminal records. The separate sections of the survey are entitled: "Current Charges," "Criminal History," "Non-Compliance," "Family Criminality," "Peers," "Substance Abuse," "Residence/Stability," "Social Environment," "Education," "Vocation," "Leisure/Recreation," "Social Isolation," "Criminal Personality," "Anger" and "Criminal Attitudes. When the scales scores are calculated, they are converted into decile scores ranging from one (lowest) to ten (highest). For a detailed discussion about COMPAS's risk-assessment process, see Freeman "Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in *State v. Loomis*" 2016 *North Carolina J L & Tech* 79–83.

40 Note "Criminal law – Sentencing guidelines – Wisconsin Supreme Court required warning before use of algorithmic risk assessments in sentencing – *S v Loomis*, 881 N.W.2d 749 (Wis. 2016)" 2017 *Harvard LR* 1531.

41 See Kehl, Guo & Kessler 11.

42 *Loomis* 754–755.

43 At 755.

> of the crime and because your history … and the risk-assessment tools that have been utilized, suggest that you're extremely high risk to re-offend."

Loomis challenged his sentence,[44] arguing that the trial court's use of the COMPAS score violated his right to due process (his constitutional right to a fair trial) – essentially that it was unfair for the court to rely on an opaque algorithm, the score of which he could not directly assess, interrogate, and challenge.[45] The Supreme Court of Wisconsin ultimately rejected Loomis's argument.[46]

3 ABILITY TO "REFUTE, SUPPLEMENT, AND EXPLAIN"

In response to Loomis's argument that a criminal defendant has a right to sentencing based on accurate information, the court acknowledged that Equivant,[47] the company that developed COMPAS, views the proprietary algorithm that generates these recidivism risk scores as a trade secret. In press interviews, the company confirmed that:[48]

> "The key to our product is the algorithms, and they're proprietary. We've created them, and we don't release them because it's certainly a core piece of our business."

The company maintains that it must shield the algorithm from scrutiny because of its proprietary nature. The result is something Kafkaesque: a criminal justice tool that does not have to explain itself.[49] Thus, neither Loomis nor his counsel was able to review or question how Loomis's score was calculated.[50]

The court reasoned that Loomis had the opportunity to "refute, supplement, and explain" the COMPAS assessment, because it was largely based on publicly available information – his answers to a 137-question survey and data about his criminal history – all of which he could verify.[51] However, there is a vast difference between the ability to verify separate pieces of information which are fed into the software, and the ability to review how the score is calculated.[52]

---

44 Loomis initially filed a motion for post-conviction relief requesting a new sentencing hearing (at 756). The court denied his motion on the basis that it would have imposed the same sentence with or without COMPAS (at 757). Subsequently, Loomis filed an appeal, and the court of appeals referred the case to the Wisconsin Supreme Court for resolution of the due process issues (*S v Loomis* 2015 WL 5446731 (Wis Ct App 17 September 2015)).

45 Citron "(Un)fairness of risk scores in criminal sentencing" *Forbes* (13 July 2016) available at https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#7a2241044ad2 (accessed on 23-06-2020). There are other significant issues raised by this case that are ripe for interrogation and analysis, for example the impact of AI systems on bias and equality in the context of criminal justice, and the impact of algorithmic sentencing tools on judicial decision-making.

46 For a comprehensive discussion of the court's treatment of all the claims in *Loomis*, see Freeman 75–106; Note 2017 *Harvard LR* 1530–1537; Liu, Lin & Chen 122–141.

47 Formerly known as Northpointe Inc.

48 Jeffrey Harmon (general manager at Equivant), as quoted in Smith *The New York Times* (2016).

49 Thompson "Should we be afraid of AI in the criminal justice system?" *The Atlantic* (20 June 2019) available at https://www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/ (accessed on 04-08-2020).

50 Pasquale comments: "[W]hen companies offer commercial rationales for keeping their 'secret sauce' out of the public eye, courts have been eager to protect the trade secrets of scoring firms" (Pasquale "Secret algorithms threaten the rule of law" *MIT Technology Review* (1 June 2017) available at https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/ (accessed on 18-06-2020)).

51 *S v Loomis* 881 NW2d 749 (Wisconsin 2016) 761.

52 Zalnieriute, Bennett Moses & Williams 22.

Neither Loomis – nor, for that matter, the court – could know to what extent COMPAS based its risk-assessment on any of the factors it claims to consider. It did not explain the breakdown of each variable, the relative weighting between variables, or the correlation between them.[53] Thus, Loomis might have seen the input and output, but he had no knowledge of their relationship. An accused is accordingly merely given an opportunity to argue against a score in the absence of any real understanding of the basis for its calculation.[54]

As commentators have pointed out, expert witnesses are subject to cross-examination, not only for their opinions, but also for their methodologies. However, the COMPAS software cannot be subpoenaed to show up to court and be questioned, although it is as powerful, if not more so, than an expert witness in influencing the court's decision.[55] Because neither the accused nor the court knows what goes into the calculation of the risk scores, the accused can, at best, present a superficial argument against the elements that may or may not be included in the algorithm.[56]

In *Townsend v Burke*,[57] the United States Supreme Court recognised that the due process right to a fair sentencing procedure included the "right to be sentenced on the basis of accurate information". The Wisconsin Court of Appeals in *S v Skaff* expounded upon this principle by underscoring that an accused must be given the "means" to investigate and verify the information:[58] "Skaff does not complain that the trial court relied on inaccurate information; he complains of the denial of means to ascertain whether there was any misinformation."

How the *Loomis* court satisfied itself of these criteria remains a mystery. The right of an accused to evaluate and assess the accuracy of information used during sentencing conflict with the very nature of the COMPAS algorithm.[59] Proprietary algorithms do not speak to ease of access. Rather they completely exclude anyone from outside the company to gain access to the source code and the way in which the scores are calculated.[60] It is abundantly clear that, without access to the source code of the algorithm, neither Loomis nor any other accused truly has the "means" to investigate and determine whether there was potentially misinformation.

It is also apparent that the majority of the court did not even have a rudimentary understanding of the operation of COMPAS. Its workings essentially consist

53 For example, in its *Practitioner Guide*, Northpointe explains the way in which it determines the Violent Recidivism Risk as follows: The Violent Recidivism Risk Scale is constructed based on characteristics including: "History of Noncompliance Scale", "Vocational Education Scale", "Current age", "Age-at-first arrest" and "History of Violence Scale," and each item is multiplied by a given weight expressed in "w", without disclosing what the "w" actually is (Liu, Lin & Chen 133).

54 Zalnieriute, Bennett Moses & Williams 22.

55 Liu, Lin & Chen 133. Frank Pasquale likens a secret algorithm that offers a damning score to an anonymous expert whom one cannot cross-examine (Pasquale (2017)).

56 Freeman 94.

57 *Townsend v Burke* 334 US 736 (1948).

58 *S v Skaff* 447 NW2d 84 (Wis Ct App 1989) 88–89. The court held that the right to be sentenced based on accurate information includes the right to review and verify information contained in the pre-sentence investigation report.

59 Freeman 88.

60 *Idem* 88–89.

of the following series of steps: (a) data input; (b) processing and computation; and (c) prediction output. It is the processing and computation step – which is the core of the COMPAS system and which involves the critical questions of how the input data is interpreted, and how the prediction output is based on those interpretations – that the *Loomis* court completely ignored. The court thus erred in holding that Loomis could challenge the accuracy of his answers to a questionnaire and his criminal history, without addressing whether and in what way he could challenge the accuracy of the processing and computation phase.[61] In its ignorance, the court prioritised business over justice.[62]

Lack of transparency was the specific focus of one of the concurring opinions in *Loomis*. Abrahamson J lamented:[63]

> "This court's lack of understanding of COMPAS was a significant problem. At oral argument, the court repeatedly questioned both the State's and defendant's counsel about how COMPAS works. Few answers were available...[M]aking a record, including a record explaining consideration of the evidence-based tools and the limitations and strengths thereof, is part of the long-standing, basic requirement that a circuit court explain its exercise of discretion at sentencing."

Such transparency and analysis of the tool itself would also, in the judge's opinion, provide "he public with a transparent and comprehensible explanation for the sentencing court's decision".[64]

Abrahamson J's concurring opinion highlights one of the critical challenges identified by both legal and technical experts. As algorithms have become an established part of high-stakes projects, there have arisen concerns that they are not adequately transparent to allow for accountability, especially if they are used as the basis for harmful or coercive decisions.[65]

4 THE "BLACK BOX" PROBLEM

In *The Black Box Society*,[66] Frank Pasquale compares the role of algorithms in the modern world to Plato's metaphor of the cave, with most people trapped and only able to see "flickering shadows cast by a fire behind them". The prisoners cannot understand the actions, not to mention the agenda, of those who create the images that are all they know of reality.[67] The problem of the "black box" is essentially that the secrecy and complexity of algorithmic processes frustrate meaningful scrutiny of automated decision-making.[68] Succinctly put: "Faulty data, invalid assumptions, and defective models can't be corrected when they are hidden."[69]

---

61 Liu, Lin & Chen 134.
62 Freeman 2016 *North Carolina J L & Tech* 95.
63 *S v Loomis* 881 NW2d 749 (Wisconsin 2016) 774.
64 At 775.
65 Sheppard "Warming up to inscrutability: How technology could change our concept of law" 2018 *U Toronto LJ* 47.
66 *The black box society: The secret algorithms that control money and information* (2015).
67 *Idem* 150.
68 Liu, Lin & Chen 134. See, generally, also Bostrom & Yudkowsky "The ethics of artificial intelligence" in Frankish & Ramsey (eds) *The Cambridge handbook of artificial intelligence* (2014) 316–334; Ormond "The ghost in the machine: The ethical risks of AI" 2020 *The Thinker* 4–10.
69 Pasquale (2015) 21.

Simon Chesterman uses the term "opacity" to denote the quality of being difficult to understand or explain.[70] He then introduces the following classification of opacity.[71] In the first instance, as in the case of the COMPAS system, opacity might arise because certain technologies are proprietary.[72] Secondly, a source of opacity might be complex systems that require specialist knowledge and skill to understand them.[73] Thirdly, there are systems that are simply naturally opaque. In machine learning, black box models are created by an algorithm directly from data. This means that humans – even those who design these systems – cannot understand how variables are being combined to make predictions.[74] As a result, the outputs of these systems are inherently less susceptible to human understanding and explanation and objective evaluation.[75]

Proprietary and complex systems can be viewed as *legal* "black box" problems.[76] The source of the opacity is the proprietary and complex characteristics of the source code, which are protected as trade secrets.[77] Private companies that develop proprietary software have both a greater interest in shrouding their products in secrecy in order to remain more competitive, and more legal tools at their disposal to keep their algorithms away from public scrutiny.[78]

However, as discussed below, *legal* "black box" issues can fairly easily be resolved by establishing a high degree of transparency and accountability, and, at the same time, protecting companies' proprietary interests. For example, the law could compel private, for-profit companies, while they perform essential public services, to disclose their algorithmic processes to the parties to litigation or court-approved panels of experts for scrutiny, and so limiting, but not obliterating, trade secret protection.[79] As Han-Wei Liu *et al* put it:[80] "Given the importance of the public interest involved in these public services, secrecy for profit should be reasonably confined."

---

70 Chesterman "Through a glass, darkly: Artificial intelligence and the problem of opacity" *NUS Law Working Paper 2020/011* (April 2020) available at http://law.nus.edu.sg/wps/ (accessed on 31-05-2020).

71 Chesterman's classification is similar to Burrell's three "forms of opacity." Burrell "'How the machine thinks': Understanding opacity in machine learning algorithms" 2016 *Big Data & Society* 1.

72 To protect an investment, detailed knowledge of the inner workings of a system might be limited to the owner of that system. Chesterman *NUS Law Working Paper 2020/011* (2020). Burrell refers to this form of opacity as "intentional secrecy", when techniques are treated as a trade or state secret. Burrell 2016 *Big Data & Society* 1.

73 These systems are nevertheless capable of being explained. Chesterman *NUS Law Working Paper 2020/011* (2020). Burrell calls this form of opacity "technical illiteracy" Burrell 2016 *Big Data & Society* 4.

74 Rudin & Radin "Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition" 1 *Harvard Data Science Review* (3 December 2019) available at https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/5 (accessed on 30-06-2020).

75 Chesterman *NUS Law Working Paper 2020/011* (2020); Zalnieriute, Bennett Moses & Williams 2019 *Modern LR* 3; Similarly, Burrell's third form of opacity relates specifically to machine learning and stems from the difficulty of understanding the actions of a complex learning technique working on large volumes of data, even equipped with the relevant expertise. Burrell 2016 *Big Data & Society* 5, 10.

76 Liu, Lin & Chen 135.

77 *Ibid.*

78 Kehl, Guo & Kessler 28.

79 Liu, Lin & Chen 135.

80 *Ibid.*

Having access to a risk-assessment algorithm's source code would enable defence attorneys to employ expert witnesses to evaluate the accused's risk scores, so that they can better attack the results during cross-examination. Probing the weaknesses of these AI systems on cross-examination could aid in exposing errors within the technology, and in so doing increase its accuracy and reliability.[81]

Far more problematic are AI systems that are opaque by design, presenting *technical* "black box" issues.[82] In the case of advanced machine learning (deep learning)[83] devices eventually outgrow their initial coding and use new sets of data to produce outcomes.[84] This implies that the calculation that led to a particular outcome is not only unknown to the consumer of the generated answer (such as the judge) but, worse, also to the designers and programmers, because the AI-enabled device has acted upon data that they are unaware of or, unbeknownst to them, it has created its own algorithms to "solve problems".[85] Put differently, "[a]s Machine Learning algorithms get smarter, they are also becoming more incomprehensible."[86] In this sense:[87]

> "[T]he fact that Machine Learning algorithms can act in ways unforeseen by their designer raises issues about the "autonomy," "decision-making", and "responsibility" capacities of AI. When something goes wrong, as it inevitably does, it can be a daunting task discovering the behavior that caused an event that is locked away inside a black box where discoverability is virtually impossible."

In sum, *technical* "black box" systems inherently lack transparency, because their decisional rules emerge automatically in ways that no-one – not even their programmers – can adequately explain.[88]

One of the central themes that both legal scholars and technology experts emphasise is the need for greater transparency about how these algorithms were developed, the assumptions that were made in their design, how the input factors are weighted, and how frequently they are assessed and updated. The challenges presented by opacity are two-fold. In the first instance, opacity makes it difficult for outside experts to evaluate and audit the algorithms in order to test for accuracy and bias.[89] Secondly, lack of information about how inputs are weighed

---

81 See Freeman 102.

82 Liu, Lin & Chen 135.

83 Deep learning occurs mainly through what are known as "artificial neural networks," which mimics the human brain. They comprise various "neurons," i.e., interconnected processors. The learning algorithms of the artificial neural network are not programmed *a priori*; rather, they "learn" the relationships between items in enormous sets of data and then formulate their own decisional rules in ways that are most often not intelligible to humans. Liu, Lin and Chen 135; Anonymous, "Artificial intelligence primer" 2018 *Victorian All-Party Parliamentary Group on Artificial Intelligence* 2 available at https://www.parliament.vic.gov.au/ images/stories/AI-Primer Feb2018.pdf (accessed on 14-03-2019).

84 Giuffrida, Lederer & Vermerys "A legal perspective on the trials and tribulations of AI: How artificial intelligence, the internet of things, smart contracts, and other technologies will affect the law" 2018 *Case Western Reserve LR* 778.

85 Knight "The dark secret at the heart of AI" *MIT Technological Review* (2017).

86 Lewis & Monett "AI & machine learning black boxes: The need for transparency and accountability" *KDNuggets* (April 2017) available at https://www.kdnuggets.com/2017/04/ai-machine-learning-black-boxes-transparency-accountability.html (accessed on 20-03-2019).

87 *Ibid.*

88 Liu, Lin & Chen 135.

89 Greater transparency will help in increasing the general understanding of these systems, how they work, and the trade-offs involved in implementing them; Kehl, Guo & Kessler 32.

means it is more difficult to bring legal challenges to the use of these tools, because the accused cannot determine how or even whether suspect factors (such as racial or gender proxies) might have influenced the risk-assessment score or the ultimate sentencing decision of the judge.[90]

Although disclosure might mitigate *legal* "black box" concerns, unfortunately, the solution to resolving *technical* "black box" issues might not be as simple as disclosing details about the data used or the source code of the algorithm that was employed. As we have seen, many of the most robust emergent machine-learning techniques are so sophisticated and opaque in their operations that they defy human scrutiny.[91] In 2012, the principal researcher at Microsoft Research New England, Tarleton Gillespie, stated: "There may be something in the end impenetrable about algorithms."[92] Others are not quite as fatalistic, but there is growing consensus among computer scientists that it would take aggressive research to cut through algorithmic opacity, particularly in machine learning, where opacity is at its densest.[93]

This opacity issue is the one that seems to be the most daunting for law-makers. Although legislation can always be passed to make a protected line of coding available for analysis in case of litigation, how does one identify how an algorithm produces an erroneous result when even its programmers cannot explain how the result was attained?[94] It is exponentially more difficult to determine what causes biased outputs in algorithms that self-program.[95] Is it the underlying data? Or is it the code that comprises the algorithm?[96]

5 "PROCESS LEGITIMACY" IN JUDICIAL DECISIONS

Chesterman points out that the legitimacy of certain decisions depends on the transparency of the decision-making process as much as they depend on the decision itself. This is best exemplified by judicial decisions.[97] More than two centuries ago, Jeremy Bentham wrote:[98]

> "Publicity is the very soul of justice.... It keeps the judge himself while trying, under trial."

Judicial decisions are the clearest example of an area in which the use of opaque AI systems should be limited.[99] The legitimacy of the rule of law depends to a large extent on whether individuals can understand the reasons for decisions affecting them, and learn how future decisions might affect them.[100]

---

90 *Idem* 28.
91 Knight "Forget killer robots – Bias is the real AI danger" *MIT Technology Review* (3 October 2017) available at https://www.technologyreview.com/2017/10/03/241956/forget-killer-robotsbias-is-the-real-ai-danger/ (accessed on 13-02-2020).
92 As quoted in Sheppard 48.
93 *Ibid.*
94 Giuffrida, Lederer & Vermerys 779.
95 Garcia "Racist in the machine: The disturbing implications of algorithmic bias" 2016 *World Policy J* 116.
96 *Ibid.*
97 Chesterman *NUS Law Working Paper 2020/011* (2020).
98 Bentham "Draught for the organization of judicial establishments" (1790) in Browning (ed) IV *The works of Jeremy Bentham* (1843) 285 316.
99 Chesterman *NUS Law Working Paper 2020/011* (2020).
100 Zalnieriute, Bennett Moses & Williams 5.

Moreover, as illustrated above, advances in computational methods, especially deep learning, come at the expense of a human's ability to explain their inferential reasoning.[101] However, accountability for legal decisions generally require that the decision-maker has a cogent reason for a particular decision or act. Reason is of particular significance in judicial decisions.[102] In the common-law tradition, the *ratio decidendi* is binding on lower courts. Appeals are usually taken on the basis of errors in the law, or in the application of the law to the facts as disclosed in the reasons.[103] In many jurisdictions, the failure to give reasons could in itself give rise to an appeal.[104]

The one matter that cannot be overlooked in the proper functioning of the legal system is the human factor. Legal issues arise out of human conduct and court decisions have an impact on the individuals who participate in them. Humans prefer that legal decisions be justified in "intelligible language, sufficiently comprehensive, and reasonably short".[105] Thus, there are semantic and pragmatic dimensions to our understanding of what makes a legal decision justified.[106] Individuals also need to feel that they are treated "fairly" in their interaction with the legal system. Fairness in this context is not only in the outcome of their case. It is the human need to be listened to.[107]

Every court case should leave the individuals engaged in it with a sense of being treated with respect, which, in turn, engenders respect for the judicial system. A law-abiding community deserves a society in which their rights and safety are respected. The role of the court is foundational to that society – software on its own will never achieve this.[108]

The sentencing decision in *Loomis* appears to be contrary to these principles. Academics and civil society roundly criticised the trial judge's reliance on COMPAS, and this reliance also became the linchpin of an appeal that almost reached the United States Supreme Court.[109] The court effectively outsourced its decision-making authority to an algorithm that is insensitive to fundamental norms of the legal system, and by so doing the court undermined its public accountability.[110]

The goal with the implementation of an AI system should never be optimisation *simpliciter*, but appropriate weighting of social and cultural norms – such as fairness, accountability, and justice – with stringent auditing to ensure that these norms are not being compromised.[111]

In administrative decisions generally, and judicial decisions in particular, the need to explain involves "process legitimacy"[112] – especially applicable in cases

101 Liu, Lin & Chen 136.
102 Chesterman *NUS Law Working Paper 2020/011* (2020).
103 *Ibid.*
104 *Ibid.*
105 Sheppard 48.
106 *Ibid.*
107 Beazley "Law in the age of algorithm" *State of the Profession Address, New South Wales Young Lawyers, Sydney* (27 September 2017) 19.
108 *Ibid.*
109 Chesterman *NUS Law Working Paper 2020/011* (2020).
110 Liu, Lin & Chen 133.
111 Chesterman *NUS Law Working Paper 2020/011* (2020).
112 *Ibid.*

where public authorities take decisions that affect the rights and obligations of individuals.[113] The inability to explain how such a decision was arrived at would, in most cases, be akin to the decision itself having been impermissibly delegated to a third party.[114] Success with regard to decisions such as these would require that AI systems be explainable and transparent. This is essential for the ability to hold the human decision-maker accountable for those decisions.[115]

During the appeal in *Loomis*, the assistant attorney-general of Wisconsin implicitly questioned whether transparency and the ability to explain were, actually, such a significant chestnut. "After all," she stated, "we don't know what's going on in a judge's head; it's a black box, too."[116] The Attorney-General of Wisconsin, Brad Schimel, was equally unperturbed. He argued that Loomis knew everything that the court knew; judges do not have access to the algorithm either.[117] The argument that the judges and the accused were equally ignorant about the inner workings of the risk-assessment software does not serve to engender the general public's faith in the criminal justice system. Is it not essential from a fairness perspective, especially given that the accused's liberty is at stake, for all parties involved to understand how the risk assessment is performed?

6 CAN "BLACK BOXES" BE MADE INTO "GLASS BOXES"?

There is a growing algorithmic accountability movement, which seeks to make the influences of these sorts of systems clearer and more widely understood.[118] Scholars and advocates have recognised the threat that automated decision-making systems pose to the rule of law.[119] Many of them argue that we can "leverage process and procedures to put guardrails around automated decision-making systems".[120]

Danielle Keats Citron advocates for the concept "technological due process", which aims to ensure that legal subjects who have been affected by algorithmic decisions have ample opportunity to challenge these decisions.[121] This is to be achieved through audit trails, education for presiding officers on machine fallibility, detailed explanations, publicly accessible code and systems testing, among other recommendations.[122]

Citron and Pasquale argue that individuals should have the "right to inspect, correct and dispute inaccurate data and to know the sources of the data".[123] Significantly, they believe that the AI systems that generate a score from said

113 *Ibid.*
114 *Ibid.*
115 *Ibid.*
116 Christine Remington as quoted in Tashea *ABA Journal* (2017).
117 Liptak *The New York Times* (2017).
118 See, for example, Diakopoulos "We need to know the algorithms the government uses to make important decisions about us" *The Conversation* (24 May 2016) available at https://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869 (accessed on 23-06-2020).
119 Waldman 622.
120 *Idem* 615.
121 See Citron "Technological due process" 2008 *Washington University LR* 1249–1313.
122 *Idem* 1305–1313.
123 Citron & Pasquale 20.

data should be public, so that each step of the process could be inspected.[124] In sum, the secrecy behind these systems has to be pierced.[125]

Pasquale also points out that there are options between "complete algorithmic secrecy" and "complete public disclosure".[126] Qualified transparency is a well-established method of enabling a panel of experts to assess protected trade secrets in order to test a system's quality, validity and reliability.

Richard Berk, a statistician and the University of Pennsylvania, believes that all companies should be required to disclose the complete content of their algorithms.[127] Berk proposes a regulatory system modelled on the way in which the United States Food and Drug Administration regulates pharmaceuticals. An algorithm developer would be required to submit the code to an agency specifically developed for this purpose for testing, similar to how prescription drugs are evaluated. The agency's process would strike a balance that would allow for public inspection of the algorithm, while protecting the developer's intellectual property.[128]

Other proposals – undergirded by process and procedure – to establish a regulatory regime for black box algorithms include: a right to explanation of automated decisions entitling an individual clarity about the process behind the model's development;[129] algorithmic impact assessments, modelled after environmental, privacy, or human rights assessments, to assess and document a system's fairness;[130] codes of conduct and whistle-blower protections to alleviate bias problems;[131] keeping humans in the loop to act as a check on automation run amok;[132] entitling data subjects to explanations about the "logic" behind an algorithmic system analogous to the General Data Protection Regulation (GDPR) of the European Union[133] [134] and the "right to be forgotten".[135]

However, all process-driven solutions have one thing in common: they emphasise the need for transparency in AI systems. However, the very concept "transparency" presupposes that AI systems are understandable and explainable. That is good and well for *legal* "black box" systems. What about *technical*

124 *Ibid.*
125 Citron *Forbes* (2016).
126 Pasquale (2017).
127 At the very least, a government entity should be created or tasked with evaluating the full contents of risk-assessment software, even if they are proprietary like COMPAS (Kehl, Guo & Kessler 32).
128 Tashea *ABA Journal* (2017).
129 Selbst & Barocas 1087.
130 Reisman *et al* "Algorithmic impact assessments: A practical framework for public agency accountability" *AI Now Institute* (2018) available at https://ainowinstitute.org/aireport-2018.pdf (accessed 31-07-2020).
131 Katyal "Private accountability in the age of artificial intelligence" 2019 *UCLA LR* 107–128.
132 Froomkin, Kerr & Pineau "When AIs outperform doctors: Confronting the challenges of a tort-induced over-reliance on machine learning" 2019 *Arizona LR* 34.
133 Regulation 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Advancement of Such Data.
134 Kaminski "The right to explanation, explained" 2019 *Berkeley Tech LJ* 199.
135 Edwards & Veale "Slave to the algorithm? Why a 'right to explanation' is probably not the remedy you are looking for" 2017 *Duke Law & Technology R* 67–80.

"black box" systems that are opaque by design so that they are not interpretable by humans?

Some researchers are intent on solving *technical* "black box" issues. They pursue "explainable AI" (also known as XAI), which can explain machine learning inferences in terms that can be understood by humans.[136] For example, Cynthia Rudin and Joanna Radin reject the widespread belief that the most accurate models for any given data science problem must be inherently uninterpretable and complicated.[137] This belief stems from the historic use of machine learning systems that were developed for low-stakes decisions, such as online advertising and web searches, where individual decisions did not deeply affect human lives.[138] Even if one has a list of the input variables, "black box" predictive models, driven by machine learning, can be such complicated functions of those variables that no human – not even their developers – can understand how the variables are jointly related to each other to reach a final prediction.[139]

Interpretable models, however, which provide a technically equivalent, but more ethical, alternative to "black box" models are different – they are *constrained* to provide a better understanding of how predictions are made.[140] Most machine learning models are not designed with interpretability constraints; they are simply designed to be accurate predictors.[141]

Rudin and Radin reject outright the belief that accuracy must be sacrificed for interpretability.[142] They argue that it is this belief that has allowed companies, such as Equivant, to market and sell proprietary and black box models for high-stakes decisions when very simple, interpretative models exist for the same task. It allows the developers to profit without considering harmful consequences to affected individuals.[143] Being asked to choose between an accurate black box and an inaccurate glass box is a false dichotomy.[144]

Various studies have shown that, in the criminal justice system, the complicated black box algorithms for predicting recidivism are not any more accurate than very simple predictive models based on age and criminal history.[145] For example,

---

136 For example, there is an XAI program at the Defence Advanced Research Projects Agency in the United States that aims to develop machine learning systems that "will have the ability to explain their rationale, characterise their strengths and weaknesses and convey an understanding of how they will behave in the future" Gunning "Explainable artificial intelligence (XAI)" *Defense Advanced Research Projects Agency Project Information* (undated) available at https://www.darpa.mil/attachments/XAIProgramUpdate.pdf (accessed on 31-07-2020).

137 The authors were part of a team that participated in the 2018 Explainable Machine Learning Challenge. The goal of the competition was to create a complicated "black box" model and then explain how it worked. Rudin and Radin's team did not follow the rules. Instead of submitting a "black box", they created a model that was fully explainable. This raised for them the question whether "black box" models are used even if they were not needed; Rudin & Radin *Harvard Data Science Review* (2019).

138 *Ibid.*

139 *Ibid.*

140 *Ibid.*

141 *Ibid.*

142 *Ibid.*

143 *Ibid.*

144 *Ibid.*

145 See, for example, Zeng, Ustun & Radin "Interpretable classification models for recidivism prediction" (2016) available at https://arxiv.org/pdf/1503.07810.pdf (accessed on 31-07-2020); Tollenaar & Van der Heijden "Which method predicts recidivism best? A

continued on next page

Angelino *et al* created an interpretable machine learning model for predicting recidivism, which considers only a few rules about someone's age and criminal history.[146] The entire machine learning model is as follows: if a person (a) has been convicted of more than three prior crimes; or (b) is 18 to 20 years old and male; or (c) is 21 to 23 years old and has been convicted of two or three prior crimes, then the person is predicted to be rearrested within two years from evaluation, and otherwise not.[147] This simple, explainable model is as accurate as the COMPAS system, which is a proprietary black box model.[148]

Rudin and Radin accordingly reject the procedural safeguards and solutions expounded upon above, because, in their attempt to mitigate the negative effects of the black box, these procedural solutions in fact extend its authority, rather than recognising that the necessity of using a black box in criminal sentencing is not a *fait accompli*.[149]

Policymakers, lawmakers, government officials, and software developers should insist that black box models for high-stakes decisions should not be used, unless it is impossible to construct an interpretable model that achieves the same level of accuracy.[150] The possibility arises that we might not have to use "black box" models for high-stakes decisions at all.

7 CONCLUSION

The improper deployment of big data and algorithms in criminal justice has every potential to undermine the right to a fair trial and transparency.[151] At the same time, however, practice irreversibly points to a global trend of increasing use of AI technology in court. For example, in early 2019, the Chief Justice of Singapore noted that "machine-assisted court adjudication" is becoming a reality.[152]

Notwithstanding the slow pace of adopting AI technology in South African legal practice, I believe that it is just a matter of time before AI innovations will make their way into the South African criminal justice system. That is because the tide has irreversibly turned in favour of the use of AI-enabled risk-assessment tools in criminal justice, including sentencing, in light of their increasingly widespread use in the United States and internationally, and because of the potential benefits they offer to overburdened criminal justice systems.[153]

---

comparison of statistical, machine learning and data mining predictive models" 2012 *J Royal Statistical Society* available at https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2012.01056.x (accessed on 31-07-2020); Angelino *et al* "Learning certifiably optimal rule lists for categorical data" 2018 *J Machine Learning Research* available at https://arxiv.org/abs/1704.01701 (accessed on 31-07-2020).

146 *Ibid.*

147 *Ibid.*

148 *Ibid.*

149 See Rudin & Radin *Harvard Data Science Review* (2019).

150 *Ibid.*

151 *Ibid.*

152 Menon CJ "Opening of the legal year" Supreme Court, Singapore (7 January 2019) available at https://www.supremecourt.gov.sg/Data/Editor/Documents/chief-justice-sundaresh-menon-address-at-the-opening-of-the-legal-year-2019.pdf (accessed on 30-05-2020).

153 Kehl, Guo & Kessler 32.

However, given the innumerable challenges, national policymakers will have to proceed very slowly and cautiously in implementing these systems.[154]

However it is achieved, the need for greater transparency about decision-making algorithms, their development, the embedded assumptions and the weighting of different variables, are crucial for the rule of law.[155] It would lead to increased understanding of these systems, the values that underlie them, and their operation, "revealing what is now obscure".[156] It would also allow affected individuals to challenge these decision-making systems.[157]

Transparency should also inform a government's decision about whether to use proprietary risk-assessment software. Since a developing country such as South Africa is much more likely to purchase these products rather than to develop tools specifically for their jurisdiction, the tension between the legitimate business interests of a private company that wants to protect its product to remain competitive, and the need for public accountability, might not be easy to resolve. It may be that the financial goals of a private company and the requirement of fairness in the criminal justice system are ultimately mutually exclusive.[158]

Ultimately, the degree of transparency in automated systems is a question of human design choices. While some methods are more difficult to render transparent, it remains the choice of the designer as to whether such methods are used in a particular system.[159] Zalnieriute *et al* point out that:[160]

> "[T]he transparency and accountability of outputs hinge on the accountability of those designing the system *for* transparency and accountability… Those designing systems should be required to design them in ways consistent with the rule of law … and be able to give an account of this has been done."

The use of COMPAS in criminal sentencing – which ultimately impacts significantly on individual liberty – is an example of a system with regard to which a high degree of transparency is needed to comply with rule of law values.[161]

At least for now, humans remain in control of governments, and they can demand explanations for decisions in natural language, not computer code.[162] Failing to do so in the criminal justice context risks ceding inherently governmental and legal functions to an "unaccountable computational elite".[163]

Angwin argues that "algorithmic accountability" entails a more sceptical approach to algorithms in general.[164] We are living in a time of general techoptimism, a time in which new technologies promise to make our lives both more efficient and enjoyable. Those technologies may help to make our justice system more equitable; or they might not. The point is we owe it to ourselves –

---

154 *Ibid.*
155 Zalnieriute, Bennett Moses & Williams 17.
156 *Ibid.*
157 *Ibid.*
158 Kehl, Guo & Kessler 33.
159 Zalnieriute, Bennett Moses & Williams 16.
160 *Ibid.*
161 *Ibid.*
162 Pasquale (2017).
163 *Ibid.*
164 As quoted in Garber "When algorithms take the stand" *The Atlantic* 30 June 2016) available at https://www.theatlantic.com/technology/archive/2016/06/when-algorithms-take-the-stand/489566/ (accessed 23-06-2020).

and to Eric Loomis and every other person whose life might be altered by an algorithm – to find out.[165]

Ultimately, humans must evaluate each decision-making process and consider what forms of automation are useful, appropriate, and consistent with the rule of law.[166] I believe that criminal sentencing should not be fully or even partly delegated to automated systems, the logic of which cannot be rendered transparent and comprehensible to accused and their lawyers.[167]

Pasquale is of the view that there is "never justification for secrecy of the algorithm" in the criminal justice system.[168] Absent adequate safety measures, losing the efficiency of algorithms is a small price to pay when an accused's right to a fair trial is at stake. No matter how useful and efficient ever-more sophisticated algorithms might be for Google searches, they are not currently – and may never be – appropriate for criminal sentencing.[169] The opportunity to be heard by an impartial adjudicator is central to the legitimacy of democratic authority.[170] There is something to be said for a sentence imposed by a human judge without the assistance of an algorithm. Judges, as humans, are not shrouded in the air of mystique and infallibility that surrounds technology. In some sense it is easier to examine and challenge their decisions when an accused suspects that bias influenced the judge's decision one way or the other,[171] because judges, for the most part, have to give reasons for the way in which they act.

As for Eric Loomis himself, he was released from Jackson Correctional Institution in August 2019, after serving his full six-year term. According to COMPAS, at least, he is at high risk to return.[172]

165 *Ibid.*
166 Zalnieriute, Bennett Moses & Williams 26.
167 *Idem* 18.
168 As quoted in Tashea (2017).
169 Freeman 106.
170 Waldman 624.
171 Freeman 106.
172 Chesterman *NUS Law Working Paper 2020/011* (2020).