Judicial Decision-Making in the Age of Artificial Intelligence

Willem H. Gravett

Abstract Artificial intelligence (AI) has become a pervasive presence in almost every aspect of society and business: from assigning credit scores to people, to identifying the best candidates for an employment position, to ranking applicants for admission to university. One of the most striking innovations in the United States criminal justice system in the last three decades has been the introduction of risk-assessment software, powered by sophisticated algorithms, to predict whether individual offenders are likely to re-offend. The focus of this contribution is on the use of these risk-assessment tools in criminal sentencing. Apart from the broader social, ethical and legal considerations, to date, not much is known about how perceptions of technology influence cognition in decision-making, particularly in the legal context. What research does demonstrate is that humans are inclined to trust algorithms as objective, and, as such, as unobjectionable. This contribution examines two phenomena in this regard: (i) the "technology effect"—the human tendency towards excessive optimism when making decisions involving technology; and (ii) "automation bias"—the phenomenon whereby judges accept the recommendations of an automated decision-making system, and cease searching for confirmatory evidence, perhaps even transferring responsibility for decision-making onto the machine.

W. H. Gravett (✉)
Department of Public and Procedural Law, Akademia, Centurion, South Africa
e-mail: willemg@akademia.ac.za

H. Sousa Antunes et al. (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, Law, Governance and Technology Series 58,
https://doi.org/10.1007/978-3-031-41264-6_15

1 Introduction

To an ever-increasing degree, Artificial Intelligence (AI) (Turing 1950, p. 433)[1] systems and the algorithms (Richie and Duffy 2018, p. 1)[2] that power them are tasked with making crucial decisions that used to be made by humans. Algorithmic decision-making based on big data (Ishwarappa and Anuradha 2015, pp. 319–320)[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 neighborhoods, and whether we are eligible for credit, healthcare, housing, education and employment opportunities, among a litany of other commercial and government decisions.

Because technological "wonders" have become so ubiquitous, because they affect our lives so profoundly, and because most of us have little understanding of how they all work, the socially constructed meaning of "technology" has become implicitly associated with optimism for what technology will bring in the future (Clark et al. 2016, p. 98). To date, not much is known about how perceptions of technology influence cognition in decision-making, particularly in the legal context. The pervasive presence of technology in almost every aspect of society and business—and its rapidly increasing pervasiveness in law—makes this a critical issue.

Classic descriptions of court processes usually emphasise the dignity, slow pace and time-honoured legal expertise of the judges and prosecutors in the criminal justice system. However, nowadays, courts have become sites where data analytics and algorithms flourish. One of the most striking innovations in the United States criminal justice system in the course of the last three decades has been the introduction of risk-assessment software, powered by sophisticated and often proprietary algorithms, to predict whether individual offenders are likely to re-offend (the so-called "risk of recidivism"). The focus of this chapter is on the latest, and perhaps most troubling, use of these risk-assessment tools: their incorporation into the criminal sentencing process.

As a general matter, automation can improve the consistency and predictability of decision-making by reducing the arbitrariness for which human decisions are

[1] 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".

[2] The term "algorithm" refers to a set of rules to be followed in calculations or other problem-solving operations, especially by a computer. In practice, "algorithm" refers to the automation of the statistical method.

[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.

well-known. Given a large number of similar questions, algorithms will provide predictable and consistent answers. Simon Chesterman states: "Whereas many evaluative decisions made by humans are based on unconscious. .. biases and intuitive reactions, algorithms follow the parameters set out for them (Chesterman 2020)."

Many scholars and practitioners view automated risk-assessment systems as a promising path toward more efficient, unbiased, and empirically-based sentencing (Hannah-Moffat 2015, p. 244). Replacing judges' discretionary decision-making with structured, quantitatively derived automated decision-making, so the argument goes, will prevent judges from "sentencing blindly", *i.e.* "over-punishing" (imprisoning offenders who present little appreciable risk to public safety) or "under-punishing" (releasing dangerous criminals into communities to re-offend) (Oleson 2011, p. 1340). Automated risk-assessment systems are frequently said to minimize both the rates and the length of incarceration for low-risk offenders, resulting in lower budgetary costs and reduced social harm (Milgram 2013; Dewan 2015). Also, predictive algorithms might save precious time for overworked judges, prosecutors and court staff (Mamalian 2011).

There is still very limited empirical research about whether automated risk-assessment algorithms actually accomplish any of these goals. There is, however, significant research that points to these automated tools leading to outcomes that are skewed because of socio-economic variables, bias in the data, and inaccurate predictions. A number of factors have been identified that render automated risk-assessment tools as potentially more akin to a Pandora's box than a panacea.

For example, algorithms are only as good as the data on which they are fed and the questions that they are asked. In practice, algorithms can reify existing disparities. These data and questions are anything but a neutral statistical exercise; practitioners are required to ask a series of directed questions about criminal histories, leisure activities, education, past criminal sentences, associates, family and relationships, emotional well-being, housing, substance abuse, family child rearing, attitudes, social assistance, finances, employment and various other issues.

Thus, it is not surprising that, generally speaking, in the United States, status as an African American would likely yield a classification of high-risk, because many African Americans live in conditions of poverty, and share "high-risk" characteristics (*e.g.* social, educational, vocational and family problems, substance abuse and histories of trauma and abuse) (Hannah-Moffat 2015, p. 245). Research has shown that race and gender are complex social constructs that cannot simply be reduced to binary variables in automated risk-assessment systems (Hannah-Moffat 2012, p. 9). Thus, automated risk-assessment systems fail to adequately control for gender or racial disparity and the potential for discriminatory outcomes. In short, automated systems sever the link between punishment and individual action, lack nuance, and miss the importance of a range of motivational, contextual, and structural factors that contribute to human action.

However, the focus of this chapter is on the potential impact that these tools might have on how judges exert their own discretion in sentencing, even if they do not themselves perceive a difference. Although these risk-assessment algorithms are not making decisions in lieu of judges (yet), it is not clear how judges should incorporate them into their decision-making processes, or how the algorithms might influence their decisions. Quantification supposedly helps to hold judges accountable and makes sentencing more consistent and efficient. The problem is, however, that little is known about the efficacy of such interventions.

As discussed below, what research does demonstrate is that humans are inclined to trust algorithms as objective, and, as such, as unobjectionable (Garber 2016). The term "automation bias" describes the phenomenon whereby judges accept the guidance or recommendations of an automated decision-making system, and cease searching for confirmatory evidence, perhaps even transferring responsibility for decision-making onto the machine (Challen et al. 2019, p. 234).

2 The Sentencing Process

Judges rightly view sentencing as a grave responsibility. In the pre-trial context, the judge's decisions are primarily binary: should the defendant stay in jail for the duration of the pre-trial period, or not? But every criminal case contains a myriad of facts, factors and features which might influence the sentence to be imposed. This general difficulty is exacerbated by the sheer number of decisions that the judge has to make in order to reach an appropriate sentence (Kehl et al. 2017, p. 14).

As a preliminary matter, the judge must determine which of the numerous facts, factors and features are relevant to the sentence, and what the appropriate weight is to attach to each. Then the judge must reach the fundamental decision about whether to remove the offender from society or choose from a litany of non-incarceration possibilities. Further decisions, among others, involve the extent of the sentence, and whether any portion thereof should be suspended, and, if so, for what period of time and under what conditions.

Moreover, the judge must not only consider the appropriate punishment for the offence, but also the risk that the offender poses, *i.e.* predicting the probability of the offender's recidivism. Historically, assessing a defendant's risk of recidivism required reliance on a judge's "intuition, instinct and sense of justice", which could result in a "more severe sentence" based on an "unspoken clinical prediction" (Hyatt et al. 2011, p. 725).

For these reasons, the allure of risk-assessment software for overburdened criminal justice systems is well neigh irresistible. The appeal of automated risk-assessment systems is that they propose to inject objectivity into a criminal justice system that has been compromised, for far too long and too many times, by human failings. Proponents of automated risk-assessment systems also claim that they make sentencing more transparent and rational (Skeem 2013, p. 300).

Automated methods are credited with giving decisions substance and making them more scientific, auditable, and, consequently, conferring the appearance of legitimacy. Support for the introduction of algorithmic risk-assessment tools then rests on the premise that they enhance professionalism by improving the defensibility and accountability of decisions, generating uniformity across regions and jurisdictions, and maintaining a perception of objective scientific validity (Hannah-Moffat 2015, p. 245).

There is no empirical evidence suggesting that a longer criminal sentence has a significant impact on a person's recidivism. Thus, it does not necessarily follow that a longer prison sentence will decrease a defendant's risk of recidivism. A judge therefore faces a more complicated question about how to use an automated risk-assessment score in sentencing, and the ultimate decision might hinge on the judge's own penological theory (Hannah-Moffat 2015, p. 245). Or, the judge might simply take a risk-averse approach and impose harsher penalties on defendant who are labeled "high-risk" by software, rather than bear the personal and societal risk of a recidivist committing another crime (Hannah-Moffat 2015, p. 245).

Clearly, judges face unique challenges to their decision-making processes in the age of AI and big data. The ramifications are well illustrated by the decision of the Wisconsin Supreme Court in S v. Loomis (2016).

3 S v Loomis

In 2013 Eric Loomis, a 31-year old black man, was arrested in La Crosse, Wisconsin, on charges related to a drive-by shooting. 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 S v. Loomis (2016, p. 754). 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 degrees sexual assault S v. Loomis (2016, p. 754). In mitigation, his attorney emphasized 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 court (trial court) ordered a pre-sentencing investigation report, which included a risk-assessment by an automated system, COMPAS, to aid the court in determining Loomis's sentence. COMPAS assessments estimate the risk of recidivism based on an interview with the defendant and information from the defendant's criminal history S v. Loomis (2016, p. 754). COMPAS assesses variables under five main areas: criminal involvement, relationships/lifestyles, personality/attitudes, family and social exclusion. The COMPAS risk assessment designated Loomis a high risk for all three types of recidivism that the system measured: pretrial recidivism, general recidivism and violent recidivism S v. Loomis (2016, pp. 754–755).

In imposing the maximum sentence of 6 years imprisonment and 5 years extended supervision, the judge specifically mentioned the COMPAS score:

> 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 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 S v. Loomis (2016, p. 755).

Loomis challenged his sentence, arguing that the trial court's use of the COMPAS score violated his right to due process, because, among other arguments, it violated his right to an individualised sentence because COMPAS relied on information about the characteristics of a larger group to calculate an inference about his personal likelihood to commit future crimes. The Supreme Court of Wisconsin ultimately rejected all of Loomis's claims.

In response to Loomis's argument about his right to an individualised sentence, the court distinguished this case case from a hypothetical one in which the risk-assessment score was either the *only* factor or the *determining* factor in a sentencing decision. In *Loomis* the automated risk score was simply one piece of information among many others that the judge considered in imposing the sentence. The court suggested that a fair trial argument might have succeeded if the risk score was the determinative or sole factor that a judge considered.

The obvious problem is that, absent a clear declaration from a judge to this effect, it is impossible to determine to what extent a judge in fact relied on an automated risk score to determine a defendant's sentence. To make matters worse, because of the operation of implicit biases, the judge herself might not know.

To ensure that sentencing judges weigh the results of automated risk-assessments appropriately, the Wisconsin Supreme Court in *Loomis* prescribed both how these assessments must be presented to trial courts, and the extent to which judges might use them. While the risk score might be useful to understand public safety considerations relating to offenders' risk reduction and management, it should not be used to determine the severity or length of the punishment, and it certainly should not constitute an official aggravating or mitigating factor in a sentencing decision. In an attempt to ensure that these limitations were adhered to, the court mandated that a judge must explain at sentencing "the factors in addition to a COMPAS risk assessment that independently support the sentence imposed" S v. Loomis (2016, p. 769).

The *Loomis* court attempted to provide a procedural safeguard to alert judges of the dangers of these assessments. The court prescribed that a "written advisement" should be included in any pre-sentencing investigation report containing a COMPAS risk-assessment score. This "written advisement of its limitations" should explain that:

1. COMPAS is a proprietary tool, which has prevented the disclosure of specific information about the weights of the factors or how risk scores are calculated;
2. COMPAS scores are based on group data, and therefore identify groups with characteristics that make them high-risk offenders, not particular high-risk individuals;
3. Several studies have suggested the COMPAS algorithm may be biased in how it classifies minority offenders;

4. COMPAS compares defendants to a national sample, but has not completed a cross-validation study for a Wisconsin population, and tools like this must be constantly monitored and updated for accuracy as populations change; and
5. COMPAS was not originally developed for use at sentencing S v. Loomis (2016, p. 770).

This "written advisement of limitations" struck a note of caution to judges about relying on the COMPAS score in a meaningful way, which was reiterated in the concurring opinions. Chief Justice Roggensack penned a separate concurring opinion to clarify that:

> [W]hile our holding today permits a sentencing court to *consider* COMPAS, we do not conclude that a sentencing court may *rely on* COMPAS for the sentence it imposes . . . [Because] the majority opinion interchangeably employs *consider* and *rely* when discussing a sentencing court's obligations and the COMPAS risk assessment tool, our decision could be mistakenly be read as permitting reliance on COMPAS S v. Loomis (2016, p. 772).

As a means to address concerns about the use of algorithmic risk-assessment tools, Justice Abrahamson also wrote separately to emphasize that, in considering these tools in sentencing, a judge "must set forth on the record a meaningful process of reasoning addressing the relevance, strengths and weaknesses of the risk assessment tool" S v. Loomis (2016, pp. 774–775).

Despite expressing concerns about the potential for unfairness and discrimination inherent in algorithmic risk-assessment tools, the Wisconsin Supreme Court nevertheless unanimously approved its use in *Loomis*. The court only superficially addressed the risks inherent in these algorithmic risk-assessment tools by adding caveats and mandating that certain disclosures accompany COMPAS scores in the pre-sentence investigation reports. However, the court was silent on the fundamental underlying question of why the scores are to be included in the risk-assessment report at all if they should not affect the length of the sentence. The court did not explain how a judge might use a defendant's risk score if she cannot change the length of the sentence based on that score (Kehl et al. 2017, p. 21).

The conclusion is warranted that the court ultimately failed to meaningfully restrict the use of these systems, in large part because it failed to consider the external and internal pressures on judges to use these automated risk-assessment tools, judges' inability to evaluate risk-assessment tools, and the effect of cognitive biases on the decision-making processes of judges.

The court's "warning label" approach—as opposed to imposing meaningful restraints—is an ineffective means of changing the ways in which judges evaluate automated risk-assessments, and it has left the door wide open for judges to be heavily influenced by the risk assessments (Liu et al. 2019, p. 130).

It is unrealistic to expect a sentencing judge, after reviewing the automated risk score, to exercise discretion without any pre-determined views of, or even bias against, the defendant. All things being equal, a high risk score will make it less likely that an offender will receive the minimum sentence or avoid incarceration (Starr 2014). Apart from the fact that no judge wants to be in a position to have to defend a lenient sentence imposed on a "high risk" defendant, especially if that defendant actually commits future crimes, the court completely ignored the "technology effect" and the role of cognitive biases supporting data reliance (specifically, so-called "automation bias" and "anchoring") on a judge's decision-making process.

4 The "Technology Effect"

Researchers have identified a tendency towards excessive optimism when making decisions involving technology (Clark et al. 2016, p. 88). Because technological breakthroughs often produce dramatic and memorable results, such as revolutionizing industries and improving our quality of life—*e.g.* smartphones, smart watches, self-driving automobiles, three dimensional printing and entertainment streaming services – such events are highly salient (Tversky and Kahneman 1974, pp. 1124–1131). By contrast, technological failures are less salient, because they neither tend to change the *status quo*, nor are they likely to be discussed in public.

The result has been that, in decision-making contexts, people develop a non-conscious or "implicit" association between technology and success through accumulated experiences in which the two are paired. This is the case, because incredible technological advancement has conditioned us to *expect* that technology would be a driver of success and progress. This bias towards optimism in technology has been labeled the "technology effect" (Clark et al. 2016, p. 88).

Once unconsciously developed, implicit associations operate quickly and automatically with regard to cognition and behavior. Chaiken's heuristic-systematic model suggests that information processing can occur along two pathways: (i) a more effortful, systematic pathway; or (ii) a more automatic (heuristic) pathway that does not involve complex information processing (Chaiken 1980, pp. 752–766). A person uses the heuristic pathway when strong cues exist about the reliability of a message, which decreases that person's motivation to engage in more effortful, systematic processing. Researchers contend that:

> [T]he . . . notion of technology has become so powerfully associated with progress and achievement, or, "success", that invoking technology in a decision context can trigger an automatic assumption that decision choices involving technology will be successful (Clark et al. 2016, p. 89).

A troubling implication in the judicial context is that there are key situational characteristics that might trigger the technology effect. For example, individuals in contexts where they are experiencing high cognitive load—such as judges experience on a daily basis—might be more susceptible to the technology effect and heuristic processing (Evans 2008, pp. 255–278).

5 “Automation Bias” and the Anchoring Effect

Beyond external pressures, judges are subject to psychological biases that encourage the use of automated risk-assessment tools. Numerous studies have shown that in courts which rely on scientific and technological tools, judges (and other individuals) are submissive to computer-generated figures and results, which might frame and condition the views of judges (Liu et al. 2019, p. 130). Individuals tend to weigh purportedly expert empirical assessments more heavily than non-empirical evidence—which might create a bias in favor of an automated risk-assessment over an offender’s own narrative. Research suggests that it is challenging and unusual for individuals to defy algorithmic recommendations (Christin et al. 2015).

For example, in a recent experiment at the Georgia Institute of Technology, a student was placed in a small office with a robot to complete an academic survey. Suddenly an alarm sounded and smoke filled the hallway outside the door. The robot, which was outfitted with a sign that read “Emergency Guide Robot”, began to move. This forced the student to make a split-second decision between escaping through the clearly marked exit through which she entered or following the robot along an unknown path and through an obscure door. Twenty six out of the 30 participants chose to follow the robot, even though it guided them away from the real exit. “We were surprised”, lead researcher, Paul Robinette, stated in an interview: “We thought that there wouldn’t be enough trust, and that [we would] have to do something to prove that the robot was trustworthy (Rutkin 2016).”

The results suggest that when people are informed that a robot (or other machine) is designed to perform a particular task, as in the case of the experiment, they will probably automatically trust it to perform that task correctly. In fact, participants in the study gave the robot the benefit of the doubt, even when the robot’s instructions were somewhat counterintuitive. In another version of the study, the majority of participants even continued to follow the robot after it appeared to have “broken down” or have frozen in place, prompting a researcher to emerge and apologize for its “poor performance” (Rutkin 2016).

Advocates of automated risk-based sentencing argue that algorithms merely provide “indicative” predictions of risk. Most judges also maintain that they do not blindly follow the results provided by the algorithm when deciding the fate of an individual offender. Rather, they claim to rely on their expertise and clinical experience to assess the offender’s personality, socio-economic situation and risk of recidivism (Rutkin 2016).

However, research in behavioral economics and cognitive psychology have shown that it is psychologically difficult and rare for any human to “override” the recommendations of an algorithm (Thaler 1999, pp. 183–206). Judges are likely to follow the predictions of an automated risk-assessment algorithm. In a survey of more than 100 Canadian judges and legal practitioners, the general perception was that automated decision-making systems were “better than clinical judgment or any form of subjective judgment . . . (Hannah-Moffat 2015, p. 244)”.

From a judge's perspective, a quantitative assessment by a software program generally seems more reliable, scientific and legitimate than almost any other source of information, including her own feelings about an offender. This is the case, not only for lay persons, but also for highly skilled professionals (Hannah-Moffat et al. 2010, pp. 391–409). It is difficult to challenge numbers and equations if you have not been trained in statistics. Thus, the danger is that when the algorithm predicts a "high" risk of recidivism, the tendency would be for judges to incarcerate, regardless of other factors.

The particular problem with the court's reasoning in *Loomis* is that it placed its trust in judges to consider the "written advisement" and evaluate the automated risk-assessment score accordingly—and this during an age in which society as a whole is heavily affected by the "technology effect". With or without a written advisement, judges consistently give technology and forensic-based evidence heavier weight than other factors, whether they consciously realise it or not (Citron 2008, p. 1271).

The impulse to follow a computer's recommendation flows from "automation bias". Studies have demonstrated that "automation bias" happens because of the tendency of most people to ascribe greater trust in the analytical capabilities of an automated system than in their own, even in the face of evidence of the systems' inaccuracies (Freeman 2016, p. 98). As noted by Danielle Citron, "[a]utomation bias effectively turns a computer program's suggested answer into a trusted final decision" (Citron 2008, p. 1272).

While automated decision-making systems have the potential to eliminate particular errors associated with human decision-making, in reality, these systems seem to merely replace these errors with new ones (Freeman 2016, p. 98). According to psychology professor Linda Sitka:

> [M]ost people will take the road of least cognitive effort, and rather than systematically analyze each decision, will use decision rules of thumb or heuristics . . . Automated decision aids may act as one of these decision-making heuristics, and be used as a replacement for more vigilant systems of monitoring or decision making (Skitka et al. 1999, p. 992).

Sitka thus views automation bias as the result of a person using an automated decision-making system as a heuristic replacement for vigilant information seeking and processing. This definition treats automation bias as similar to other biases and heuristics in human decision-making (such as, for example, confirmation bias), except that automation bias stems specifically from interaction with an automated system.

Three main factors have been assumed to contribute to automation bias. First, there is the tendency of humans to choose the road of least cognitive effort (the so-called "cognitive miser hypothesis"). Thus, humans tend to use directives or recommendations of automated systems as a strong decision-making heuristic in the place of effortful cognitive processes of information analysis and evaluation (Parasuraman and Manzey 2010, p. 392).

A second factor is humans' perceived trust of automated systems as powerful agents with superior analytical capabilities. As a consequence, humans might overestimate the performance of an automated decision-making system and might ascribe to the automated systems greater capability and authority than in themselves or other humans (Dzindolet et al. 2002, pp. 72–94).

A third contributing factor to automation bias is the phenomenon of "diffusion of responsibility" (Parasuraman and Manzey 2010, p. 392). When humans share decision-making tasks with machines, the same psychological effect occurs when humans share tasks with other humans, *i.e.* so-called "social loafing", which is reflected in humans' tendency to reduce their own effort when working within a group, as opposed to when they work individually on a task (Karau and Williams 1993, pp. 681–706). To the extent that human users see an automated decision-making system as another team member, the humans might believe themselves to be less responsible for the outcome, and, as a result, reduce their own effort in monitoring and analysing other available data.

Also, because individuals and agencies often turn to algorithms with the express purpose to reduce human bias and error, these algorithms could be seen as authoritative sources, with more knowledge than the humans who interpret them. Thus, the human users, such as judges, tend to adhere to what the algorithm decides, despite the fact that such adherence might harm others. This is because of the general power that authority figures hold and "people's willingness to conform to the demand of . . . authority" (Skitka et al. 1999, pp. 992–993). Automation bias is a robust phenomenon that renders the "written advisements" mandated by the Wisconsin Supreme Court inane.

The Wisconsin Supreme Court in *Loomis* expressed its expectation that "the circuit court [would] exercise discretion when assessing a COMPAS risk score with respect to individual defendant". As explained above, the court's expectation has no basis. Human beings trust computer-generated decisions far more than they should. In fact, they rely on automated decisions even when they suspect malfunction.

A related problem is that automated decision-making systems might also "provide cover for human agents" (Chesterman 2020). For example, a survey of judges and lawyers in Canada found that many regarded software, such as COMPAS, as an improvement over subjective human judgment (Hannah-Moffat 2015, pp. 244–247). Although these practitioners did not deem risk-assessment software as particularly reliable predictors of future behavior, they nevertheless favored these systems because using them minimized the risk that the judges and lawyers themselves would be blamed for the consequences of their decisions (Hannah-Moffat 2015, p. 244). As Chesterman rightly notes, automated risk-assessment systems:

> [S]hould not be the basis for avoiding accountability in the narrow sense of being obliged to give an account of a decision, even if after the fact, or to avoid responsibility for harm as a result of that decision (Chesterman 2020).

Apart from automation bias, courts in the United States have repeatedly recognized that cautionary statements do little to prevent a factfinder from considering certain factors once the fact-finder's consciousness has been exposed to those factors. The court in *United States v Rodriguez* explained why jury instructions to disregard a personal opinion expressed by a prosecutor are not effective: "one cannot unring a bell"; "after the thrust of a saber it is difficult to say forget the wound"; and "if you throw a skunk into the jury box, you can't instruct the jury not to smell it" (United States v Rodriguez 1978).

The concern is that judges might adapt their sentencing practices in order to match the predictions of risk-assessment algorithms. Behavioral economists refer to "anchoring" to describe the common phenomenon according to which individuals draw upon any available piece of evidence—regardless of how weak it is—to make subsequent decisions (Tversky and Kahneman 1974, pp. 1128–1130; Mussweiler and Strack 2000, p. 495).

In their classic experiment on the "anchoring effect", Amos Tversky and Daniel Kahneman gerrymandered a wheel of fortune marked from 0 to 100 to stop at either the number 10 or 65. They would stand in front of a group of University of Oregon students they recruited as participants, spin the wheel, and ask the students to write down the number on which the wheel stopped (which of course was either 10 or 65). Then the experimenters asked the participants the following question: "Is the percentage of African nations among UN members larger or smaller than the number you just wrote down?"

As Kahneman explains:

> The spin of a wheel of fortune – even one that is not rigged – cannot possibly yield useful information about anything, and the participants. . . should simply have ignored it. But they did not ignore it (Kahneman 2011, p. 119).

When the wheel landed on 10, the participants provided a mean estimate of 25%; when the wheel landed on 65, the participants provided a mean estimate of 45%. Thus, simply being presented with a number—even one that they knew was totally random and which had no bearing whatsoever on the quantity they had been asked to estimate—had a pronounced impact on the participants' responses.

Since Tversky and Kahneman's seminal study, the anchoring effect has been shown to be a "truly ubiquitous phenomenon that has been observed in a broad array of different judgment domains" (Mussweiler et al. 2000, p. 1143). It has proven to be a robust, reliable, and persistent cognitive bias (Wilson et al. 1996, pp. 387–402; Mussweiler 2002, pp. 67–72). Many findings indicate that clearly irrelevant numbers—even if they are blatantly determined at random—may guide numeric judgments that are generated under conditions of uncertainty (Chapman and Johnson 1999, pp. 115–153).

It should come as no surprise that judges are not immune to anchoring effects. Judicial decisions often involve quantification. And judicial quantification generally occurs under circumstances that are inherently uncertain. First, judges must make their decisions at least partially on the basis of controverted and contradictory evidence. Second, judges are often called upon to quantify the unquantifiable—the qualitative misdeeds of the guilty party—which must be expressed as the award of monetary damages or the determination of criminal fines or length of imprisonment. In the absence of strict, algorithmic guidelines or other institutional specifications, this process can be both ambiguous and extremely subjective.

Thus, if the risk-assessment algorithm's prediction of the offender's risk of recidivism is higher than that upon which the judge settled in her own mind, she might increase the sentence, even without being consciously aware that she is following the algorithm. As stated above, it is impossible to determine to what extent a judge in fact relied on an automated risk score to determine a defendant's sentence. A judge presented with an assessment that reveals a higher risk of recidivism than predicted, may increase a defendant's sentence without realizing that anchoring might have played a role in the judgment.

6 Conclusion

The data scientist, Cathy O'Neil, describes the current age as one of unquestioned techno-optimism (Van Hollebeke 2016). The faith we tend to put in the power of technology shields algorithmic systems from critical interrogation in general. It is ironic, notes the investigative technology journalist, Julia Angwin, that we—eas a criminal justice system, political body and culture—take an all-too human approach to algorithmic infrastructure:

> We trust it too much. We have not yet thought as rigorously or as strategically as we need to about its effects. We have not fully considered whether, and indeed how, to regulate the algorithms that are . . . regulat[ing] our lives . . . (Garber 2016).

Loomis is significant because it demonstrates, not only the challenges that courts face in understanding how automated risk-assessment systems work, but also the fact that there is virtually no precedent to guide judges' decision-making in using these and other automated tools.

Mere written warnings do not seem to be able to satisfactorily inform judges as effective gatekeepers, especially when they might not be sufficiently equipped with knowledge and understanding about how these automated tools function. Although warnings might alert judges to the inadequacies of these tools, the advisement might nevertheless fail to negate the considerable external and internal pressures of a criminal justice system championing the use of automated quantitative risk-assessments.

The Wisconsin Supreme Court in *Loomis* seemed impercipient to this reality. It accepted on face value the circuit court's claim in post-conviction proceedings that it "would have imposed the exact same sentence" even without the automated risk score (S v. Loomis 2016, p. 771). The court's required advisement suggests that judges should be a bias check on a tool itself designed to correct judges' biases.

A useful starting point might be to reflect anew on the question: "Who is the decision-maker?" As Liu et al. (2019, p. 138) note: "In a world that often blindly portrays numbers to be scientific, neutral and objective, human decision-makers are likely to surrender their powers to data."

As seen in *Loomis*, ill-informed deference to algorithms marginalizes the role of public authority and scrutiny in governance. As "words yield to numbers" in criminal sentencing (Hamilton 2015, p. 6), the judiciary should exercise considerable caution in assessing the qualitative value of these new technologies. Although governments might mandate humans to make final decisions, it remains problematic to address the anchoring effect, as long as data-driven approaches to decision-making processes in the public sector are tolerated.

Judges must be trained about the phenomenon of automation bias. Studies have shown that individuals who receive such training are more likely to scrutinise an automated system's suggestions (Citron 2016).

Beyond training, the best ways to ensure fairness of the automated scoring system is through procedural safeguards. In the development of algorithms for deployment in the criminal justice system, accountability and oversight are key. Policymakers must ensure that these systems have been designed for the purpose for which they are used, and that they are continually monitored and assessed for accuracy and reliability.

Facilitating outside research and auditing to evaluate and test algorithms for bias is also of critical importance. The design, implementation and evaluation of automated systems to be used in the criminal justice system should be consistent with the core values of such a system, including equal protection and due process. Important normative and ethical questions loom large at every turn as these algorithmic risk-assessment tools are integrated into the existing system – and those decisions should not be made lightly or with insufficient information.

At least for now, humans remain in control of governments, and they can demand explanations for decisions in natural language, not computer code. Failing to do so in the criminal justice context risks ceding inherently governmental and legal functions to an "unaccountable computational elite" (Pasquale 2017). Criminal justice policy should be informed by data, but we cannot afford to allow the sterile language of science to obscure questions of fairness, accountability and justice (Starr 2014).

Angwin argues that "algorithmic accountability" entails a more skeptical approach to algorithms in general (Garber 2016). We are living in a time of general tech-optimism, a time in which new technologies promise to make our lives both more efficient and enjoyable. Those technologies may help to make out justice system more equitable; they might not. The point is we owe it to ourselves—and to Eric Loomis and every other person whose life might be altered by an algorithm—to find out (Garber 2016). Ultimately, humans must evaluate each decision-making process and consider what forms of automation are useful, appropriate and consistent with the rule of law.

In the final analysis, 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 a judge's decisions when a defendant suspects that bias influenced the judge's decision one way or the other, 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 (Chesterman 2020). [4]

References

Chaiken S (1980) Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J Pers Soc Psychol 39:752–766

Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K (2019) Artificial intelligence, bias and clinical safety. BMJ Qual Saf 28:231–237

Chapman GB, Johnson EJ (1999) Anchoring, activation, and the construction of values. Organ Behav Hum Decis Process 79:115–153

Chesterman S (2020) Through a glass, darkly: artificial intelligence and the problem of opacity, NUS law working paper 2020/011. http://law.nus.edu.sg/wps/. Accessed 31 May 2020

Christin A, Rosenblat A, Boyd D (2015) Courts and predictive algorithms. Data Soc. https://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf. Accessed 30 Oct 2020

Citron D (2016) (Un)fairness of risk scores in criminal sentencing. Forbes https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#7a2241044ad2. Accessed 23 June 2020

Citron DK (2008) Technological due process. Wash Univ Law Rev 85:1249–1313

Clark BB, Robert C, Hampton SA (2016) The technology effect: how perceptions of technology drive excessive optimism. J Bus Psychol 31:87–102

Dewan S (2015) Judges replacing conjecture with formula for bail. The New York Times. https://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html. Accessed 15 Oct 2020

Dzindolet MT, Pierce LG, Beck HP, Dawe LA (2002) The perceived utility of human and automated aids in a visual detection task. Hum Factors 44:79–94

Evans JS (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 59:255–278

[4] *See* generally, on biases, in this book P G Marques - AI Instruments for Risk of Recidi-vism Prediction and the Possibility of Criminal Adjudication Deprived of Person-al Moral Recognition Standards – Sparse Notes from a Layman; and D Durães, P M Freitas and P Novais - The Relevance of Deepfakes in the Administration of Criminal Justice. *See* also, on the different applications of Machine Learning and AI, in this book A Oliveira and M A T Figueiredo - Artificial intelligence - historical context and state of the art; I Trancoso, N Mamede, B Martins, H S Pinto and R Ribeiro - The impact of language technologies in the legal domain; J Gonçalves-Sá and F L Pinheiro - Societal Implications of Recommendation Systems - A Technical Perspective; A T Freitas - Data-driven approaches in healthcare - challenges and emerging trends; M Correia and L Rodrigues - Security and Privacy; E Magrani and P G F Silva - The Ethical and Legal Challenges of Recommender Systems Driven by Artificial Intelligence; M Lanz and S Mijic - Risks associated with the use of natural language generation - Swiss civil liability law perspective; M S Fernandes and J R Goldim - Artificial Intelligence and Decision Making in Health - Risks and Opportunities; W Gravett - Judicial Decision-making in the Age of Artificial Intelligence; and D Durães, P M Freitas and P Novais - The Relevance of Deepfakes in the Administration of Criminal Justice. *See* finally, on the COMPAS system, in this book P G Marques - AI Instruments for Risk of Recidivism Prediction and the Possibility of Criminal Adjudication Deprived of Personal Moral Recognition Standards – Sparse Notes from a Layman.

Freeman K (2016) Algorithmic injustice: how the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. N C J Law Technol 18:75–106

Garber M (2016) When algorithms take the stand. The Atlantic. https://www.theatlantic.com/technology/archive/2016/06/when-algorithms-take-the-stand/489566/. Accessed 23 June 2020

Hamilton M (2015) Adventures in risk: predicting violent and sexual recidivism in sentencing law. Ariz State Law J 47:1–57

Hannah-Moffat K (2012) Actuarial sentencing: an “unsettled” proposition. Justice Q 30:270–296

Hannah-Moffat K (2015) Partiality, transparency, and just decisions the uncertainties of risk assessment. Fed Sentenc Rep 27:244–247

Hannah-Moffat K, Maurutto P, Turnbull S (2010) Negotiated risk: actuarial illusions and discretion in probation. Can J Law Soc 24:391–409

Hyatt JM, Chanenson SL, Bergstrom MH (2011) Reform in motion: the promise and perils of incorporating risk assessments and cost-benefit analysis into Pennsylvania sentencing. Duquesne Law Rev 49:707–749

Ishwarappa K, Anuradha J (2015) A brief introduction on big data 5Vs characteristics and hadoop technology. Procedia Comput Sci 48:319–324

Kahneman D (2011) Thinking: fast and slow. Farrar Straus & Giroux, New York

Karau SJ, Williams KD (1993) Social loafing: a meta-analytic review and theoretical integration. J Pers Soc Psychol 65:681–706

Kehl D, Guo P, Kessler S (2017) 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, Cambridge

Liu HW, Lin CF, Chen YJ (2019) Beyond State v Loomis: artificial intelligence, government algorithmization and accountability. Int J Law Inf Technol 27:122–141

Mamalian CA (2011) State of the science of pretrial risk assessment. Pretrial Justice Institute, Bureau of Justice Assistance, Washington, DC

Milgram A (2013) Why smart statistics are the key to fighting crime. TED, New York

Mussweiler T (2002) The malleability of anchoring effects. Exp Psychol 49:67–72

Mussweiler T, Strack F (2000) Numeric judgments under uncertainty: the role of knowledge in anchoring. J Exp Soc Psychol 36:495–518

Mussweiler T, Strack F, Pfeiffer T (2000) Overcoming the inevitable anchoring effect: considering the opposite compensates for selective accessibility. Personal Soc Psychol Bull 26:1142–1150

Oleson JC (2011) Risk in sentencing: constitutionally suspect variables and evidence-based sentencing. South Methodist Univ Law Rev 64:1329

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52:381–410

Pasquale F (2017) Secret algorithms threaten the rule of law. M.I.T. Technology Review. https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/. Accessed 18 June 2018

Richie DR, Duffy JD (2018). Artificial intelligence in the legal field. In: Association of corporate counsel greater Philadelphia in-house counsel conference

Rutkin A (2016) People will follow a robot in an emergency – even if it’s wrong. New Scientist. https://www.newscientist.com/article/2078945-people-will-follow-a-robot-in-an-emergency-even-if-its-wrong/. Accessed 20 Oct 2020

S v. Loomis (2016) 881 N.W.2d 749 (Wisc. 2016)

Skeem J (2013) Risk technology in sentencing: testing the promises and perils (commentary on hannah-moffat, 2011). Justice Q 30:297–303

Skitka LJ, Mosier KL, Burdick M (1999) Does automation bias decision-making? Int J Hum Comput Stud 51:991–1006

Starr S (2014) Sentencing by the numbers. The New York Times. https://www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html. Accessed 15 Sept 2020

Thaler RH (1999) Mental accounting matters. J Behav Decis Mak 12:183–206

Turing AM (1950) Computing machinery and intelligence. Mind 236:433–460

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131

United States v Rodriguez (1978) 585 F.2d 1234 (5th Cir. 1978)

Van Hollebeke M (2016) Shining a light on the darkness. Data Soc https://points.datasociety.net/shining-a-light-on-the-darkness-432adf10c7a0. Accessed 23 June 2020

Wilson TD, Houston CE, Etling KM, Brekke N (1996) A new look at anchoring effects: basic anchoring and its antecedents. J Exp Psychol Gen 125:387–402