John Zerilli. 2020. "Algorithmic Sentencing: Drawing Lessons from Human Factors Research," in *Principled Sentencing and Artificial Intelligence*, ed. J. Ryberg, J. Roberts & J. de Keijser. New York: OUP.

Algorithmic Sentencing: Drawing Lessons from Human Factors Research

John Zerilli

Abstract: Researchers in the field of "human factors" have long been aware that when humans devolve certain of their functions to technology, the transfer from human to machine can restructure more than the division of labour between them: humans' perceptions of themselves and their abilities may also change. In particular, if a system becomes reliable enough, humans will become diffident to the point of adhering to the system's recommendations even when they have grounds to disbelieve them. Such findings are relevant to the use of algorithmic and data-driven technologies, but whether they hold up in the specific context of recidivism risk assessment is only beginning to be considered. In this chapter, I describe and analyse some pertinent human factors results, and assess the extent to which they pose a problem for the use of algorithms in the sentencing of offenders. While the findings from human factors research are themselves robust, they do not seem to translate neatly to the judicial sphere. The incentives, objectives, and ideologies of sentencing judges appear to upset the usual pattern of results seen in many other domains of human factors research.

Keywords: automation bias, automation complacency, human factors, human-computer interaction, sentencing

The use of forecasting techniques in sentencing makes most sense on a variety of consequentialist assumptions about the aims of punishment. If the purpose of punishment is to prevent crime—the most obvious consequentialist justification for state-inflicted suffering—then an accurate assessment of an offender's likelihood of reoffending will be necessary for establishing whether (and if so to what extent) the offender should be incapacitated (e.g. by a term of imprisonment). Likewise, efforts to deter or rehabilitate an offender presumably stand to benefit from a calibrated response to the offender's unique propensity to reoffend.[1] It is thus against the backdrop of such consequentialist aims that the uptake of data-driven risk assessment has taken place.

Next to consequentialism, however, lies a more immediate source of inspiration for the use of data-driven methods in sentencing, namely, evidence that actuarial procedures for assessing risk are generally more reliable than unaided clinical (or "professional") judgment (Meehl 1954; Dawes et al. 1989). The higher accuracy—or rather *perceived* accuracy—of actuarial procedures, in turn, raises an important issue concerning their adoption, particularly in their machine-learning/big-data guise. The subfield of cognitive psychology known as "human factors" is relevant here.

"Human factors" is concerned with the psychological and ergonomic aspects of human-machine interaction. Its exploration of principles of optimal interface design and task allocation is both experimental and applied—theories of how human cognitive and physiological constraints bear upon human interaction with machines are tested empirically, and the findings of these empirical investigations are then regularly fed back into the design of human-machine systems. The aim of human factors researchers is to

discover principles that enable humans to interact with technology in the safest, most productive, error-minimising and comfortable way. In the usual domains in which such investigations have occurred (marine transportation, aviation and a number of industrial and manufacturing arenas), the most persistent findings are striking. It appears that once a system reaches a particular threshold of accuracy and reliability, its human invigilators are likely to fall into a state of complacency or deferred criticism: the system, in effect, is assumed to be all-knowing, so that outputs which might otherwise be viewed with suspicion by a diligent and conscientious observer tend to be overlooked or excused on the assumption that “the machine knows best” (see Parasuraman & Manzey 2010 for reviews). So understood, the phenomenon is not unlike the so-called “CSI effect,” in which police, jurors and even judges are liable to overestimate the importance of forensic evidence (Marks et al. 2017; see also Damaška 1997).

The consequences of human factors research are potentially far-reaching for sentencing policy, though the field of human factors itself has had very little to say about the use of technology in the legal profession. For as automated decision aids become ever more widespread in the sentencing of offenders, it is natural to worry that judges will fall prey to the same tendencies to which other professionals, technicians and experts have succumbed when using state-of-the-art software tools (Cummings 2004). It was just this worry to which a recent French report on artificial intelligence gave expression when it noted that “it is far easier for a judge to follow the recommendations of an algorithm which presents a prisoner as a danger to society than to look at the details of the prisoner’s record himself and ultimately decide to free him” (Villani 2018, 124). The New York-based AI

Now Institute expressed fears along similar lines: “[w]hen [a] risk assessment [system] produces a high-risk score, that score changes the sentencing outcome and can remove probation from the menu of sentencing options the judge is willing to consider” (AI Now 2018, 13).

The question remains, however, whether we really should be worried by all this. In this chapter, I describe and analyse some pertinent human factors results and assess the extent to which they pose a serious problem for the use of algorithms in the sentencing of offenders. While the findings from human factors research are themselves robust, they do not seem to translate neatly to the judicial sphere. The incentives, objectives, and ideologies of sentencing judges appear to upset the usual pattern of results seen in many other domains of human factors research.

The field of human factors

Human-computer interaction has been studied since at least the late 1960s (Pazouki et al. 2018; see Kelley 1968, Edwards and Lees 1974, and Sheridan and Ferrell 1974, for early reviews). While initially the topics receiving greatest attention concerned optimal task allocation, interface design and software ergonomics (Rouse 1981; Hatvany and Guedj 1982; Williges and Williges 1982), these fairly niche research projects can be seen to form part of a broader preoccupation with the psychology of computer-aided decision-making. Still, it was not until the publication of Lisanne Bainbridge’s (1983) paper on the “Ironies of Automation” that the most pressing psychological challenges of computer-aided

decision-making were diagnosed.[2] In her work, the contradictions inherent in computer-aided decision-making were made explicit, the principal one being "that the more advanced a control system is, so the more crucial may be the contribution of the human operator" (1983, 775). By this she meant that because no system can ever quite be fail-safe, a human must always play at least some invigilatory or oversight role (rendered all the more crucial precisely because the system is so advanced); but because automation gives rise to the kinds of cognitive phenomena we have mentioned, it may be that no human *can* adequately perform an invigilatory role (Bainbridge 1983, 776).

In other work (Zerilli et al. 2019; Zerilli et al. 2021), I have subdivided the overarching human factors problem into four distinct subproblems, which can also be seen as four distinct (but closely allied) areas of human factors research (see Box 1). Roughly, the introduction of automation can lead to: a *capacity problem* when automated systems compute over many more variables than human cognitive limits allow the human to process or keep pace with; an *attentional problem* when the task at hand involves little more than monitoring a system's generally seamless transactions, making it very difficult for the human to sustain visual attention and maintain proper "situation awareness"; a *deskilling problem* when human skills are not adequately maintained through regular exercise, and consequently degrade; and finally an *attitudinal problem* when a system reaches such a level of proficiency as to induce overtrust in its human overseer. These problems may occur on their own or together in any combination. They may also reinforce one another in particular cases (e.g. the first three problems could easily exacerbate the last one).

BOX 1

Four problems investigated in the field of human factors

1. The capacity problem

Humans are not able to keep track of the systems they are tasked with supervising because the systems are too advanced and operate at incredible speeds.

2. The attentional problem

Humans get very bored very quickly if all they have to do is monitor a display of largely static information.

3. The deskilling problem

Skills that are not regularly maintained will degrade over time (also known as "deskilling").

4. The attitudinal problem

Humans have a tendency to overtrust systems that perform reliably *most* of the time (even if they are not reliable *all* of the time).

Adapted from Zerilli et al. (2021, 85)

The capacity problem is a potentially serious one for the law. In many of the most socially consequential and normatively loaded applications of machine learning (such as risk assessment), knowing that an automated system works properly *just is* knowing its reasons for deciding, and to be satisfied with those reasons. A judge relying on a risk assessment tool, for example, would want to know that the system reasoned legitimately to

its conclusion, and did not (let us say) take an offender's race, religious beliefs or sexual orientation into account when assigning a risk score. In fact, without some such understanding of the system's operations, it is hard to see how a risk score could inform a judge's discretion in a rational way. How, for instance, would a score enter into a judge's calculations without the judge knowing what importance to place upon the score, as determined (among other things) by the quality of the reasoning that led to it? A judge's being unable to understand why a system assigns the scores it does (as the capacity problem implies could be the case) seems to pose a genuine ethical and legal conundrum.

Be that as it may, I shall not pursue the capacity problem any further in this chapter. For one thing, the explanation of machine learning decisions is a topic that interests researchers well beyond human factors, and which in fairness merits separate treatment.[3] For another, much of the interest of the capacity problem from a human factors point of view lies in its exacerbation of the attitudinal problem.[4] Indeed, to the extent that both problems arise in a particular case, the capacity problem could make some efforts at alleviating the attitudinal problem somewhat futile. If I am tempted to overtrust a system that is very complex—indeed, overtrust it partly because it is so complex—exhorting me *not* to overtrust it, and use my better judgment, is unlikely to jolt me out of my dependence if my dependence results partly from my inability to understand how the system works.

The deskilling problem, too, is likely to have important ramifications for sentencing practice, inasmuch as being less exercised in traditional (algorithmically unassisted) sentencing tasks could over time lead to sentencing judges becoming gradually less adept

in the kind of practical reasoning required for sentencing deliberation. Nevertheless, sentencing is not wholly unlike other forms of judicial and forensic deliberation which today are still conducted manually (so to speak) and at which judges can therefore be expected to maintain their skills—e.g. when making findings of credibility, or setting levels of aggravated and punitive damages in civil trials. Moreover, the exact nature of and potential for this sort of “moral deskilling” remain to be clarified, both as a philosophical and empirical matter (see Vallor 2015). Thus I shall not pursue the deskilling problem any further here. And given that risk scores are not “monitored” in any relevant sense of the word, the attentional problem does not really arise. So for the remainder of this chapter I shall restrict my analysis to the attitudinal problem alone.

Bad attitudes: Automation complacency and automation bias

Two manifestations of the attitudinal problem have received significant attention over the past few decades (see e.g. Skitka et al. 2000; Parasuraman and Manzey 2010; Pazouki et al. 2018). *Automation-induced complacency* describes conditions in which, owing to the highly automated nature of a task, the human operator’s role has ceased to *actively* involve them, so that they are no longer impelled to assess a system’s outputs critically, and lapse into an unduly diffident, deferential or unsuspecting state of mind with respect to the system (Pazouki et al. 2018). *Automation bias* occurs when human operators preference a system’s signals over other sources of information, including the evidence of their own senses (Pazouki et al. 2018). These two closely related phenomena “describe a conscious

or unconscious response of the human operator induced by overtrust in the proper function of an automated system” (Parasuraman & Manzey 2010, 406). Disturbingly, they seem obstinately resistant to intervention. There is evidence that explicit briefings about the risks associated with the use of a particular tool are not enough to counteract the strength of automation bias, and that extended practice likewise is ineffective against automation complacency (Parasuraman & Manzey 2010).

Intriguingly—and fortunately—these problems only appear to arise within a fairly narrow band or “sweet spot” of system performance. When a system is not regarded as especially reliable, the effects are not seen (Bagheri and Jamieson 2004; Parasuraman & Manzey 2010; Banks et al. 2018a; Banks et al. 2018b). The effects are seen only when reliable *but imperfect* systems are used, so that automation is considered “most dangerous when it behaves in a consistent and reliable manner *for most of the time*” (Banks et al. 2018b, 283, emphasis added). On the other hand, when a system functions reliably *all* of the time, or at any rate more reliably than its human counterpart (by some margin), the effects *are* seen, but arguably do not matter, or matter less in proportion as the system exceeds human performance (Zerilli et al. 2019, Zerilli et al. 2021). Hence the sweet spot: the problem arises at a certain threshold of system performance, but wanes (at least in terms of the danger it poses, if not its existence) once the system measurably outperforms an expert human counterpart.

How worried should we be about all this when it comes to the use of risk assessment tools in sentencing? If these results were to generalise to the judiciary, I submit that we should be very worried. First, on a typical construal of the judicial function in

sentencing, risk assessment is not meant to be delegated to an algorithm, any more than a judge can delegate the task of fact-finding to an expert witness. Despite the expert being more knowledgeable about a particular province of learning, the judicial fact-finding role is still one which the judge is personally expected to discharge. In a similar vein, automated risk assessment tools are intended to serve in the literal sense of automated decision *support* tools, so that the decision-maker exercises their own independent judgment, ideally before consulting the tool. On this picture, the tool functions as little more than a check on the decision-maker's intuitions, and should in no way be seen as a substitute for the decision-maker's actual discretion. This is a crucial point. There are in fact many ways that an algorithm could be said to "support" a human decision-maker (Administrative Review Council 2004, 14-15, 20). Extended along a scale of types of support, these would range from least instructive to most instructive—from merely scaffolding, prompting or perhaps supplementing human judgment at one end, to more actively coaxing or even replacing the human decision-maker at the other end. I have previously suggested that risk assessment tools can be situated on the "less instructive" end of this scale, as systems that *supplement* human judgment by carrying out functions in such a way as to augment human capacities (Zerilli et al. 2019). So if it turns out that the risk of a judge succumbing to automation complacency is real, we should all be worried, since it amounts to no less than the risk of a vital judicial responsibility being abdicated.

Second, the credentials of risk assessment instruments locate them firmly in the problematic "sweet spot" zone I mentioned earlier—the zone where their perceived reliability and accuracy vis-à-vis human decision-makers is liable to induce overtrust in

them. As I mentioned earlier, statistically-guided methods of decision-making have been shown to have a certain edge over unstructured clinical/professional judgment. But this superiority is not unqualified. Statistical methods and data-driven algorithms may be reliable, but they are not unequivocally reliable—i.e. significantly better-than-human in a preponderance of hard cases. The main reason for this is that not all relevant factors to a determination will be codified into an algorithm, and even when they are, the determination itself (such as whether someone poses a risk of recidivism) may be only one component of a larger decision (such as regarding what sentence to impose). Marion Oswald (2018, 16), when addressing the use of algorithms in the public service, rightly warns that assuming “that the forecast or classification represents the only or main factor on which the ‘rightness’ or ‘wrongness’ of the overall decision is to be judged…may risk changing the question that the public sector decision-maker has to answer.” While her remarks were made with public sector decision-makers in mind, they are no less relevant to sentencing judges. Sentencing never reduces to brute prediction. No jurisdiction is so fixed on a single objective. Thus even if the human “gets it wrong” (i.e. mispredicts) while the machine “gets it right” (e.g. the offender *did* end up reoffending, though the judge let them out), the fact remains that there will always be more to sentencing than a sole consideration, and the judge always obliged to take stock of factors not feasible (or even possible) to encode in an algorithm (such as retributive considerations, general deterrent considerations, remorse/contrition, and the like). Proper decision-making will have regard to all of these factors, even if the outcome *looks* wrong. (This is to underscore a point I made earlier when discussing the capacity problem: reasons for a decision matter, and *how*

a judge (or a machine) decides is arguably even more important than *what* the judge (or a machine) decides.) The use of algorithms in sentencing may obscure this point, because algorithms do have a certain edge over human case workers where predictions are concerned, and it is this very superiority that can make a foil of human judges. The conditions, we could say, are ripe for automation complacency and bias.

The upshot of the foregoing discussion, then, is that yes, *if* the human factors results were to hold in the judicial sphere, we should indeed be concerned. But now the question is, *do* those results, in fact, hold? The limited research record suggests that the results do not hold; and it is worthwhile inquiring into why this might be the case.

Preliminary investigations

Angèle Christin's (2017) ethnographic study compared the use of algorithms by web journalists and legal professionals, including court staff and judges. She notes that "the discussion [regarding the uptake of algorithmic instruments] has largely focused on the instruments themselves," and that "[w]e know less about the practices…of the people who rely on algorithmic technologies in their work and lives" (Christin 2017, 2). Her work takes place against a backdrop of previous investigations into sociotechnical systems (e.g. Orlikowski 1992; 2007) that prise open the organisational contexts in which technological artefacts operate. An important aim of this work is to expose how an artefact's meaning (and therefore reception) within an organisational setting may be actively shaped by the organisation's practices, policies, chains of command, and the like. Such work arguably

provides grounds for skepticism about the impact that risk assessment instruments are likely to have in the criminal justice system, for it is by no means a given that the techno-utopian rhetoric that often surrounds the arrival of a new technology will prove justified in this organisational setting in particular.

While Christin did not directly set out to address the salience of human factors in the criminal justice system, the questions she asked are certainly pertinent here: "How do people make sense of the recommendations provided by algorithmic tools? Do they blindly follow the algorithms' suggestions, manipulate the instruments, or ignore them? How do algorithmic practices and representations vary depending on their context?" (Christin 2017, 2). Her methodology involved *in situ* observation of criminal proceedings in three courts in the US, as well as interviewing 22 court personnel, including court administrators, probation officers, judges and defence lawyers. She found discrepancies between managerial claims and the actual day-to-day use of algorithms by those on the ground ("During misdemeanor and felony hearings, most judges and prosecutors did not use the analytics, dashboards, and risk assessment tools at their disposal"). She quotes one judge as saying:

> I don't look at the numbers. There are things you can't quantify…You can take the same case, with the same defendant, the same criminal record, the same judge, the same attorney, the same prosecutor, and get two different decisions in different courts. Or you can take the same case, with the same defendant, the same judge,

> etc., at a two-week interval and have completely different decision [*sic*.]. Is that justice? I think it is. (Christin 2017, 9)

She discusses a variety of strategies used by court staff to minimize the impact of algorithms. One of them (so-called "foot-dragging"), involves simply "ignoring or bypassing risk scores" altogether (Christin 2017, 9). Another strategy involves conscious gaming of the system to achieve a desired output. Overall, despite a great deal of hype and ongoing controversy regarding their various biases, Christin found that little attention is paid to algorithms in practice.

Much the same conclusions were reached by Megan Stevenson's (2018) study of the US state of Kentucky's experiment with several pretrial risk assessment tools since 2011. She notes that while Kentucky's statutes, via a clear set of "action directives," strongly favoured pretrial release as the default setting, these directives were not followed. Had they been followed, 90 per cent of defendants would have been granted "immediate non-financial release" (Stevenson 2018, 311). In fact only 29 per cent were granted such terms at the first bail setting. As she observes: "If judges are not convinced or coerced to follow statutory guidelines, a risk assessment tool will not be an effective method of liberalizing release" (2018, 311).

Garrett and Monahan's (2020) study reports *inter alia* on a set of investigations into the US state of Virginia's experience with the use of non-violent risk assessment (NVRA) instruments in sentencing. Very much in line with Christin and Stevenson's findings, their surveys of Virginian judges reveal "highly divergent attitudes towards risk assessment"

and that "[a] sizable minority of judges [have] great discomfort with the goals and the use of risk assessment at sentencing" (Garrett and Monahan 2020, 445). Eight out of ten judges endorsed the view that sentencing should be based on more than just the gravity of the offence committed and so should factor in the offender's risk of reoffending. But by implication, two out of ten judges did not agree with the use of risk assessment in sentencing, and perhaps endorse purely retributivist sentencing aims: either way "a significant minority of judges excluded considerations of risk when sentencing eligible drug and property offenders and were largely unfamiliar with the NVRA" (Garrett and Monahan 2020, 468). This is consistent with around half the judges stating that they "always" or "almost always" considered NVRA results in drug and property matters, and around a third stating that they "usually" did so in such cases (Garrett and Monahan 2020, 467). A further interesting result is that seven out of ten judges believed the availability of noncustodial and rehabilitative options to be "less than adequate," and 75 per cent considered that the availability of more options would "change their sentencing practices" (Garrett and Monahan 2020, 467).

A final study (Grgić-Hlača et al. 2019) examined the effects of algorithms on bail determinations, using civilian volunteers instead of judges to test a range of hypotheses. While the use of risk assessment instruments for bail determinations might not be immediately relevant to sentencing, the study's primary results are still worth citing. In general, the authors report that receiving machine advice has only a small effect on decisions ("For most cases, the fraction of participants who predict recidivism is very similar with and without advice") (Grgić-Hlača et al. 2019, 8). Where participants make

their predictions before receiving machine advice, and their predictions diverge from the machine’s, in only a minority of cases (19.9 per cent) will they change their pre-advice predictions so as to concur with the machine’s advice after learning of it. This minority (290 out of 5150 cases) is a subset of a somewhat larger minority of cases (390 out of 5150) in which participants change their predictions in response to machine advice (thus in 100 of 5150 cases participants changed their predictions but not their advice). Furthermore, participants are not more likely to follow machine advice when they are given (positive) information about its accuracy. On the other hand, *incentivising* participants to follow machine advice does seem to make them more sensitive to the advice (although interestingly “the effect is not more pronounced if the incentive is stronger”) (Grgić-Hlača et al. 2019, 16).

The role of incentives, objectives and ideology in sentencing practice

Why might the judicial use of algorithms fail to conform to patterns of algorithmic use in other domains of activity which human factors researchers have investigated?

Perhaps the fact that sentencing judges must by law exercise their own judgment in sentencing, with many judges taking this to mean that they should do so *before* consulting the algorithm, could be upsetting the usual pattern of results. The bail study provides modest evidence that this could be the case, for in only a minority of cases where participants made their predictions *before* receiving machine advice did they then go on to change their predictions *after* receiving that advice (and fewer still actually changed their

advice). What makes this interpretation a little less compelling, however, is the fact that participants were told in advance that the machine had an accuracy rate of 68 per cent, and might therefore have “read this as a hint not to take the advice seriously” (Grgić-Hlača et al. 2019, 9). In fact, the bail study could be interpreted as offering solid evidence for another hypothesis entirely: that warnings to judges about the relatively low accuracy of a risk assessment tool are effective in counteracting some of the effects of automation complacency and bias. If so, while there would still be a discrepancy between what is observed in the courts and what has been observed in other forums, the discrepancy would also be easy enough to explain in human factors terms: automation complacency and bias do not occur when an autonomous system is perceived to be only moderately reliable. Telling judges that a tool is accurate in fewer than 7 cases out 10 may be enough to knock them out of their complacency. Perhaps the most effective warning would give a quantitative indication of accuracy *as well as* an instruction to use the tool merely as a check on one’s working *after* one has already made up one’s mind. This is a worthwhile proposal that should be the subject of future criminological/human factors inquiry. (Of course, to the extent that one remains in the dark about how an algorithm arrives at its conclusions—even in general terms that cite the various factors in a decision and their weights—using the algorithm as “check” on one’s work still seems rather like a stab in the dark.)

Yet another possibility is that there simply *is* no discrepancy to speak of here: judges *are* being influenced by risk assessment instruments, perhaps in salutary ways too—e.g. in their desire to reduce incarceration rates—it is just that they are constrained

from giving effect to algorithmic recommendations by a lack of adequate resources. Garrett and Monahan (2020) did find, after all, that a large majority of judges were potentially willing to follow NVRA advice if only the interventions it recommended were actionable. A plausible twist on this scenario might posit that when judges already have reason to expect that it will not be possible to follow through with an algorithm’s recommendations (say, because of resource constraints), they will systematically discount or ignore them.

But a few other hypotheses have better empirical support. The first emphasises the likely crucial role that the *incentives* of judges play in structuring their patterns of algorithmic reliance and aversion. Human factors scholars have long known that accountability mechanisms can be effective in offsetting tendencies to automation bias. As the authors of one experiment put it, “making participants accountable for either their overall performance or their decision accuracy led to lower rates of automation bias” (Skitka et al. 2000, 701). What is implicit in this result is that when participants are given prudential reasons to use instruments carefully, they are more likely to do so. To some extent this is also borne out by the bail study, which found that incentives to find ground truth, or to avoid false positives or false negatives, do not induce reliance on machine advice. Fair enough, then: users of algorithms can be induced to handle algorithmic information more rationally and discriminatingly when given prudential reasons to do so. And perhaps sentencing judges have enough of these prudential motivations operating in the background to explain why they do not easily succumb to automation bias. But while incentives are almost certainly operating in the background of sentencing deliberations,

they do not seem to be the kinds of incentives that necessarily lead to more careful sifting of algorithmic information.

Firstly, incentives can be perverse. The bail study is a case in point. In addition to finding that participants could be incentivised to find ground truth in spite of machine advice, it also revealed that participants could be incentivised to rely on machine advice. I interpret these results to mean that incentives can be effective in any direction: where the incentives are targeted to finding ground truth, *that* behaviour is what the incentives are likely to elicit, not reliance on machine advice (especially when machine advice is distrusted); when the incentives are instead geared toward reliance on machine advice, *that* behaviour is what the incentives are likely to elicit (to some extent even if machine advice is distrusted); and so on.

Secondly, there is evidence that sentencing judges can indeed be incentivised in less than ideal ways—they may be prudentially motivated to follow or to ignore algorithmic recommendations in ways that do not track their perceptions of an algorithm’s reliability. (In other words, there is evidence that sentencing judges are not wholly motivated to find ground truth.) More concretely, incentive schemes may encourage judges to discount algorithmic outputs *even when they consider the outputs to be reliable*, and, conversely, to follow algorithmic outputs *even when they consider the outputs to be unreliable*. The following is by no means a far-fetched scenario (see e.g. Stevenson and Doleac 2019, 20). A judge might (consciously or otherwise) rely on an algorithm to take some of the pressure off them for releasing a “low risk” offender that they believe has a very slight chance of reoffending (say, a likelihood just better than chance). In such a case,

the judge technically disagrees with the algorithm (which assigns “low risk”), but acquiesces to the recommendation regardless, because the judge is willing to give the offender the benefit of the doubt and can cite the algorithm in defence in the event that the offender recidivates. It is true that the judge here does not *strongly* disagree with the algorithm, and could in some sense be said to be following its recommendation. The point, however, is that the algorithm has tipped the scales of justice in favour of lenience partly because it offers the judge an excuse for a poor outcome in a case that could have gone either way (Stevenson and Doleac 2019, 5, 20; see also Van Dam 2019). At all events, the situation is not one where we can say that the judge “uncritically” followed the algorithm: the judge would personally prefer to ignore its recommendation, but chooses to heed it for partly self-interested reasons. In the converse situation, a judge may bypass an algorithm’s designation of an offender as “low risk” out of fear of a public backlash and a need to be (seen to be) tough on crime—even though the judge personally agrees with the algorithm and would otherwise release the offender. In these and similar ways, incentives may structure patterns of algorithmic dependence and deviation in ways that do not track judicial perceptions of an algorithm’s reliability.

Another set of incentives may make judges less disposed to place their trust in algorithms overall. Christin (2017, 11) observes that “the long training process and high barriers to entry in the field of law shape the professional identity of judges and prosecutors in powerful ways, making them more likely to doubt the benefits of using external tools to complement or replace their own expertise.” That professional incentives could operate in these ways should not be surprising. Incentives are woven into the fabric

of the legal profession. It is known that career advancement and professional esteem are strong motivators on the bench (Shepherd 2011; Cooter 1983). In jurisdictions where elevation to judicial office depends on being elected, such motivations are even more plain (Brace et al. 1999). And once appointed (or elected), judges obviously strive to avoid being overturned on appeal (Randazzo 2008). It is therefore not surprising that these incentives would mesh in complex ways with external pressures to use algorithms in sentencing, and potentially render the latter less effective as a result.

Apart from operating under unique incentive structures, judges are plausibly less prone to automation complacency and bias because they also have unique *objectives*. A judge may think that a risk score is reliable enough, but also happen to think that risk scores should not inform sentencing—e.g. in accordance with a retributive theory of punishment (see Garrett and Monahan 2020, 445, 468). Such a judge will not set crime prevention as a sentencing objective. They may, for example, view a young offender as being less culpable, and consequently impose a lighter sentence, despite youth being a high predictor of recidivism and an algorithm assigning a high risk score. But even when a judge is not a retributivist, and simply takes account of more than just an offender's recidivism risk in sentencing (as indeed they must), the effect of such competing objectives will likely mitigate the influence of an algorithmic assessment, even if it will not eliminate that influence entirely (Stevenson and Doleac 2019).

Finally, judges bring unique *ideologies* to sentencing. Ideologies overlap with incentives and objectives, but can be singled out too as uniquely influencing judicial behaviour. As Christin (2017, 10) notes once again: "In criminal justice, innovation does

not come with the glitter and appeal that it has in other sectors: it is often a source of uncertainty, because by definition an innovation arrives without the vetting of precedent.” One might say that the ideology of the legal profession as a whole is against innovation, quantitative analysis, and forecasting—its vehicles of reasoning (precedent) and redress (compensation, restitution, retribution, etc.) are predominantly backwards-looking. Legal culture is nothing if not steeped in tradition. This ideology plausibly frames an *a priori* suspicion of algorithmic techniques in criminal justice that manifests as algorithmic aversion and the belief (rightly or wrongly) that risk assessment instruments are less-than-reliable. This is another way of saying that algorithmic aversion may itself be an ideology, to which the tradition-steeped ideology of the legal profession naturally leads.

One factor I have not mentioned is judicial bias, which could well be lumped in with ideology. Judges are obviously not above holding biases against certain demographics, and there is emerging evidence that algorithms, far from mitigating their effects, can give them licence (Albright 2019). Importantly, this phenomenon is not always due the algorithms themselves being biased (which is a separate issue) (Stevenson 2018, 309). Entrenched stereotypes may affect the way a judge will interpret the same risk score assigned to two offenders differing only in their socioeconomic status, so that the disparity cannot really be attributed to the risk score. Skeem et al. (2019) found that judges who, without a risk assessment to hand, might have been more lenient on relatively poor offenders than more affluent ones, may impose harsher penalties on poor offenders—and lighter penalties on more affluent offenders—the moment a risk assessment tool forms part of the sentencing calculus. This is not because risk assessment tools are necessarily biased

against poor offenders. Indeed all the offenders who were assigned risk scores in the study were assigned exactly the same risk scores, regardless of their socioeconomic status.

It is an arresting result, which apparently held even after controlling for the gender, race, political orientation and jurisdiction of the judges. The study's authors speculate that it arises from the difference between assessments of blameworthiness and assessments of risk. They reason that low socioeconomic status often plays an exculpatory role in sentencing, in contrast to affluence. But when attention is diverted from the assessment of blameworthiness to the assessment of *risk*—as it inevitably is under risk assessment—low socioeconomic status becomes a disadvantage to the extent that it is prejudicially perceived as indicative of higher recidivism risk. The perception can probably aptly be described as "prejudicial" here because affluent offenders assigned the *same* risk score did not receive penalties that were as harsh as those visited on the poorer ones. In the authors' own words (references omitted, emphasis added):

> Adding formal risk assessment information may have cued judges to process poverty as a factor that increased the likelihood that the defendant would continue committing offenses…This context may have activated stereotypes of poverty and affluence that led judges to interpret *identical risk scores* as signaling a much higher risk of rearrest for the relatively poor defendant than his more affluent counterpart." (Skeem et al. 2019, 57)

Conclusion

In this chapter, I have described and analysed some pertinent human factors results, and assessed the extent to which they pose a problem for the use of algorithms at sentencing. I conclude that while the findings from human factors research are robust, they cannot be applied straightforwardly to sentencing judges. The incentives, objectives, and ideologies of judges may exert a significant gravitational pull away from algorithmic sentence recommendations or risk predictions, so that judges are unlikely to blindly accept what a machine tells them about an offender's risk of recidivism. Judicial incentives and ideologies in particular may be such as to make judges less prone to the allure of data-driven and high-tech innovation.

To be sure, algorithmic sentencing may pose genuine challenges—challenges of transparency, bias, data protection, and so on. It can also be expected to pose many of the same challenges as those posed by the human processing of statistical information more generally, such as the various heuristics and biases discussed by psychologists and behavioural economists (e.g. the availability heuristic, anchoring bias, overconfidence, etc.). However, the fear that judges will fall victim to *automation* bias, unthinkingly parroting whatever an algorithm happens to say in a spirit of "Computer says NO," is not one that preliminary evidence suggests is well-founded.

I have, in passing, suggested that appropriately crafted warnings may have something going for them in any event. It is true that a warning to judges not to place too much stock in an algorithm would not tell them *how* to discount it, and, for reasons I

explained, discounting becomes virtually impossible to do rationally when a judge does not understand how a given algorithmic assessment is calculated. Nevertheless, there are some early signs that warnings may be enough to dispel any illusions judges might have about a technology, sufficient to encourage them to consult risk assessments only *after* they have come to their own conclusions. While this would not resolve the discounting issue, it could be enough to mitigate automation-induced complacency and bias. Further research should investigate this matter directly.

References

Administrative Review Council. 2004. *Automated Assistance in Administrative Decision Making*. Barton, ACT: Commonwealth of Australia.

AI Now. 2018. *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems*. New York: AI Now Institute.

Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." John M. Olin Center for Law, Economics, and Business Fellows' Discussion Paper Series No. 85. Available at: http://www.law.harvard.edu/programs/olin_center/fellows_papers/pdf/Albright_85.pdf

Bagheri, N., and G.A. Jamieson. 2004. "Considering Subjective Trust and Monitoring Behavior in Assessing Automation-Induced 'Complacency.'" In *Human Performance, Situation Awareness, and Automation: Current Research and Trends*, edited by D. A. Vicenzi, M. Mouloua, and O. A. Hancock, pp. 54-59. Mahwah, NJ: Erlbaum.

Bainbridge, Lisanne. 1983. "Ironies of Automation." *Automatica 19* (6): pp. 775-779.

Banks, V.A., A. Erikssona, J. O'Donoghue, and N.A. Stanton. 2018a. "Is Partially Automated Driving a Bad Idea? Observations from an On-Road Study." *Applied Ergonomics 68*: pp. 138-145.

Banks, V.A., K.L. Plant, and N.A. Stanton. 2018b. "Driver Error or Designer Error: Using the Perceptual Cycle Model to Explore the Circumstances Surrounding the Fatal Tesla Crash on 7th May 2016." *Safety Science 108*: 278-285.

Brace, Paul, Melinda Gann Hall, and Laura Langer. 1999. "Judicial Choices and the Politics of Abortion: Institutions, Context, and the Autonomy of Courts." *Albany Law Review 62* (4): pp. 1265-1302.

Christin, Angèle. 2017. "Algorithms in Practice: Comparing Web Journalism and Criminal Justice." *Big Data and Society 4* (2): pp. 1-14.

Cooter, Robert D. 1983. "The Objectives of Private and Public Judges." *Public Choice 41* (1): pp. 107-132.

Cummings, M.L. 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems." *AIAA 1st Intelligent Systems Technical Conf*. (https://doi.org/10.2514/6.2004-6313).

Damaška, M.R. 1997. *Evidence Law Adrift*. New Haven: Yale University Press.

Dawes, R.M., D. Faust and P.E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science 243* (4899): pp. 1668-1674.

Edwards, E., and F.P. Lees, editors. 1974. *The Human Operator in Process Control*. London: Taylor and Francis.

Garrett, Brandon L., and John Monahan. 2020. "Judging Risk." *California Law Review 108*: pp. 439-493.

Grgić-Hlača, Nina, Christoph Engel, and Krishna P. Gummadi. 2019. "Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing." *Proc. ACM Human-Computer Interaction 3* (CSCW 178): pp. 1-25.

Hatvany, J., and R.A. Guedj. 1982. "Man-Machine Interaction in Computer-Aided Design Systems." Proc. IFAC/IFIP/IFORS/IEA Conf. *Analysis, Design and Evaluation of Man-Machine Systems*, Baden-Baden, Sept. Oxford: Pergamon Press.

Kelley, C.R. 1968. *Manual and Automatic Control*. New York: Wiley.

Marks, A., B. Bowling, and C. Keenan. 2017. "Automated Justice? Technology, Crime, and Social Control." In *The Oxford Handbook of Law, Regulation, and Technology*, edited by Roger Brownsword, Eloise Scotford, and Karen Yeung, pp. 705-730. New York: Oxford University Press.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press.

Orlikowski, Wanda J. 1992 "The Duality of Technology: Rethinking the Concept of Technology in Organizations." *Organization Science 3* (3): pp. 398-427.

Orlikowski, Wanda J. 2007. "Sociomaterial Practices: Exploring Technology at Work." *Organization Studies 28* (9): pp. 1435-1448.

Oswald, Marion. 2018. "Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power." *Philosophical Transactions of the Royal Society A 376*: pp. 1-20.

Parasuraman, R., and D.H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors 52* (3): pp. 381-410.

Pazouki, K., N. Forbes, R.A. Norman, and M.D. Woodward. 2018. "Investigation on the Impact of Human-Automation Interaction in Maritime Operations." *Ocean Engineering* 153: pp. 297-304.

Randazzo, Kirk A. 2008. "Strategic Anticipation and the Hierarchy of Justice in US District Courts." *American Politics Research 36* (5): pp. 669-693.

Rouse, W.B. 1981. "Human-Computer Interaction in the Control of Dynamic Systems." *ACM Computing Surveys 13*: pp. 71-99.

Shepherd, Joanna. 2011. "Measuring Maximizing Judges: Empirical Legal Studies, Public Choice Theory and Judicial Behavior." *University of Illinois Law Review 2011* (5): pp. 1753-1756.

Sheridan, T.B., and W.R. Ferrell. 1974. *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*. Cambridge, MA: MIT Press.

Skeem, Jennifer L., Nicholas Scurich, and John Monahan. 2019. "Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants." *Law and Human Behavior 44* (1): pp. 51-59.

Skitka, L. J., K.L. Mosier, and M. Burdick. 2000. "Accountability and Automation Bias." *International Journal of Human-Computer Studies 52*: pp. 701-717.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review 103*: pp. 303-384.

Stevenson, Megan T., and Jennifer L. Doleac. 2019. "Algorithmic Risk Assessment in the Hands of Humans." Available at: https://ssrn.com/abstract=3489440 or http://dx.doi.org/10.2139/ssrn.3489440

Vallor, Shannon. 2015. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy and Technology 28*: pp. 107-124.

Van Dam, Andrew. 2019. "Algorithms Were Supposed to Make Virginia Judges Fairer. What Happened was Far More Complicated." *Washington Post*, November 19.

Villani, C. 2018. *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. Available at: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

von Hirsch, Andrew. 1976. *Doing Justice*. New York: Hill and Wang.

von Hirsch, Andrew. 1986. *Past or Future Crimes*. New Brunswick, NJ: Rutgers University Press.

Wickens, C.D., and C. Kessel. 1979. "The Effect of Participatory Mode and Task Workload on the Detection of Dynamic System Failures." *IEEE Trans. Syst., Man, Cybern. 9* (1): pp. 24-31.

Wiener, E.L., and R.E. Curry. 1980. "Flight-Deck Automation: Promises and Problems." *Ergonomics 23* (10): pp. 995-1011.

Williges, R.C., and B.H. Williges. 1982. "Human-Computer Dialogue Design Considerations." Proc. IFAC/IFIP/IFORS/IEA Conf. *Analysis, Design and Evaluation of Man-Machine Systems*, Baden-Baden, Sept. Oxford: Pergamon Press.

Zerilli, John, John Danaher, James Maclaurin, Colin Gavaghan, Alistair Knott, Joy Liddicoat, and Merel Noorman. 2021. *A Citizen’s Guide to Artificial Intelligence*. Cambridge, MA: MIT Press.

Zerilli, John. Forthcoming. “Explaining Machine Learning Decisions.” *Philosophy of Science*.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. “Algorithmic Decision-Making and the Control Problem.” *Minds and Machines 29* (4): pp. 555-578.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2018. “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” *Philosophy and Technology 32* (4): pp. 661-683.

---

[1] For penological arguments against the use of predictive instruments, and more generally against preventative aims in sentencing, see von Hirsch (1976, ch.3; 1986, 176-178).

[2] For precursors to Bainbridge, see Wickens and Kessel (1979) and Wiener and Curry (1980).

[3] See Ryberg, and Chiao, this volume; Zerilli (forthcoming) (for a philosophical account of the problem and a framework within which to approach its resolution); and Zerilli et al. (2018) (for an earlier attempt at the same).

[4] Cf. Ryberg, this volume.