Reference: Eijk, G. van (2021). Algorithmic Reasoning: The Production of Subjectivity Through Data. In: R. Peeters & M. Schuilenburg (Eds.), *The Algorithmic Society. Technology, Power, and Knowledge*. London: Routledge (pp. 119-134). *Please note that the published text may be slightly different.*

Algorithmic reasoning: The production of subjectivity through data

Gwen van Eijk, Erasmus University Rotterdam

Introduction: dehumanizing algorithms?

In a letter to *InsideTime*, the English National Newspaper for Prisoners and Detainees, former prisoner Charles Hanson writes about the algorithmic tool OASys (Offender Assessment System), describing it as a 'dehumanising process of risk assessment'. Since 2001, OASys informs parole and probation decisions and case management (Robinson 2017). Hanson situates his comments within a broader critique of rehabilitation programmes, but takes aim particularly with the role of OASys in how 'prisoners are seen as "things" to be measured, assessed, quantified and computerised' which 'has tended to dehumanise the offender':

> 'Taking groups of offender's human identities apart and categorising them into axis, tables, graphs and risk assessments by the use of computers and assuming that this is scientific is an act of intellectual dishonesty which distorts what makes up the human condition and the potential for change' (Hanson 2009).

Hanson's letter indicates how algorithmic practices have transformed criminal justice, in this case probation and rehabilitation practices. His letter touches on various long-standing concerns about predictive justice. First, predictive justice conflicts with individual justice – the notion that each case should be assessed on its own merits (Binns 2019). Algorithmic predictions are not individualized but rather assess the extent to which an individual shares certain characteristics with a group of individuals that is known to reoffend (Hannah-Moffat 2013). Predictive sentencing has therefore been criticized because it 'would be legally and morally wrong to take action against a person based on membership in a specific class or other group status' (Simmons 2018, 1076). According to Eckhouse and colleagues (2019, 189), the decision to use data on groups to make decisions about individuals is the 'base layer' of bias that is fundamental to other layers of bias (such as biased data or biased models). This is specifically problematic when risk assessment includes information about socioeconomic status, which produces class, racial and gender bias in criminal-legal decisions (cf. Starr 2014; Hannah-Moffat 2016; van Eijk 2017).

Because of concerns about bias, various commentators, including myself, have concluded that the use of predictive algorithms cannot be legitimized in sentencing decision but have left the door open for legitimate use in treatment decisions (van Eijk 2017; in press; Hamilton 2015; Tonry 2019a). This is not to say that there are in the context of rehabilitation no concerns about algorithmic justice. Hanson rightfully expresses concern that OASys 'distorts what makes up the human condition and the potential for change' which touches on the question of human dignity in algorithmic justice. From a human rights perspective, Ward (2011, 106) has suggested that labelling individuals as 'high-risk' could violate the moral principle of human dignity, as it could impact 'the degree to which a person is free to form his or her own intentions and is able to act in accordance with them without interference'. This raises the question whether algorithmic rehabilitation practices are compatible with the core idea of rehabilitation, that is, that desistance requires *agency:* a sense of having control to change one's course in life

(McNeill 2006; Healy 2013). An important aspect of transformative agency is that individuals can envision an alternative future self (Paternoster and Bushway 2009; King 2012). The prediction of future behaviour may conflict with envisioning a positive future self, which may even be reinforced by algorithmic prediction.

In this Chapter I will argue that the problem of algorithmic justice in the context of risk-based rehabilitation is not in the first place a technological problem but a problem of human-algorithm interaction and the rationalization of algorithms. Interpretative errors seem inherent to what algorithmic predictions aim to do, namely, predict an individual's future behaviour based on group data. In this way, prediction conflicts with the principle of individual justice, which may subsequently result in violating the principle of respect for human dignity. Prediction of future behaviour and the way in which prediction are decontextualized from structural factors may discourage individuals to re-envision a positive future self which may hamper agency and, consequently, desistance. In the conclusion I offer several principles for improving ethical use of algorithms in the context of rehabilitation.

Human-algorithmic practices: rehabilitation, risk and marginality

The use of algorithmic predictions has proliferated and algorithms now inform decision-making processes at an increasing number of stages of the criminal justice system, from pretrial detention to release (for an overview and discussion of different uses for different decisions see: van Eijk 2020). To give an idea: in the U.S., twenty-eight states and several counties use risk assessment tools for sentencing decisions (Stevenson and Doleac 2018), while at the Federal level algorithmic risk assessment has a role in allocating treatment and resources (Monahan and Skeem 2016; Bussert 2019). In addition, an increasing number of states and counties is using risk assessment for pretrial decisions, which in many states is introduced as an alternative to the bail system (Desmarais and Lowder, 2019). In Canada, Hong Kong and European jurisdictions among which the U.K., the Netherlands and Finland, algorithmic tools inform presentencing reports and treatment plans in the context of rehabilitation (Hannah-Moffat 2013; Robinson 2017; Salo et al. 2015; van Wingerden et al. 2014).

To worry about human decision-makers being replaced by machines is however unfounded, as humans continue to be involved in criminal-legal decisions. Rather, algorithmic tools *inform* decisions made by judges, parole boards and probation officers about individuals (Green and Chen 2019). Hannah-Moffat, Maurutto and Turnbull (2009) have even spoken of an 'actuarial illusion' as they saw that Canadian correctional workers engaged with predictive tools on their own terms: they may overrule predictions, devalue predictive tools and prefer to rely on their own expertise (see also van Wingerden et al. 2014; Werth, 2017). Binns (2019, 19) similarly notes that 'screen-level bureaucrats' dealing with algorithmic systems exercise discretion according to their own commitments which may conflict with the goals of their organisation. Based on observations, Hannah-Moffat *et al.* (2009) contend that risk is best understood as a negotiated process: practitioners simultaneously embrace and resist risk technologies.

However, keeping humans 'in the loop' of algorithmic decision-making, as Hanson as well as other critics suggest (e.g. Binns 2019; Simmons 2018; La Diega 2018), is no guarantee for justice. One reason for this is that we should understand risk-based

decision-making as the outcome of the *interplay* of humans and algorithms. Therefore, to understand the uses and consequences of predictive justice it is essential to investigate the human-algorithm interaction and to not reduce one aspect to the other (Winner 1980; Binns 2019). Indeed, we need not only ask how we can improve algorithms but look at how to improve professional engagement with algorithmic predictions (Green and Chen 2019).

A central question of an ethics of algorithms is 'What do algorithms do to subjects?' (Matzner 2017, 28). This concerns not only the question of how algorithms shape decisions, but also how algorithmic reasoning shapes the production of the 'criminal subject': they shape how practitioners view justice-involved individuals as moral subjects (cf. Hannah-Moffat 2005; Werth 2019). Such insight is particularly important when algorithms inform rehabilitation practices that require convicted individuals to reflect on their behaviour and to transform from a 'high-risk offender' to a law-abiding individual. I started this Chapter with Hanson's letter because the literature on algorithmic justice tells us surprisingly little about how justice-involved individuals who are subjected to algorithmic practices experience it. Simmons (2018) has theorised that algorithmic decisions may violate the principle of procedural justice, which implies a negative experience for those who are subjected to algorithms. Indeed, in the context of loans, insurance, hiring and other day-to-day situations, Binns and colleagues (2018) observe that algorithmic decision-making is considered unfair by people subjected to it, as it 'reduces a human being to a percentage.' Studying algorithmic and human decisions in managerial decisions that require human skills (e.g. hiring and work evaluation), Lee (2018) found that people perceive human decisions as fairer and more trustworthy than algorithmic decisions, and that people expressed more negative emotions in response to algorithmic decisions. Decisions in the criminal justice system pre-eminently involve human skills, which suggests that involving algorithms in these decisions is likely to be negatively experienced by those subjected to them.

To fully understand what algorithms 'do' to people, we need to study them in the context of criminal-legal policies and practices in which they are shaped and used. In this Chapter, I focus on rehabilitation practices, specifically the use of predictive tools in the context of risk-based justice. The use of algorithmic tools is embedded in a currently dominant approach to rehabilitation: the Risk-Needs-Responsivity (RNR) model developed in the 1990s by Canadian psychologists Andrews and Bonta (2010a; see also Hannah-Moffat 2013; Taxman et al. 2014). The RNR model is used in Northern American and European jurisdictions and follows several principles. The risk principle prescribes *who* should be treated: direct intensive services to the higher risk offenders and minimize services to the low risk offenders, the needs principle prescribes *what* should be treated: criminogenic needs, while the responsivity principle addresses *how* the intervention should be delivered: characteristics of the intervention are important as well as individual 'strengths, ability, motivation, personality, and bio-demographic characteristics such as gender, ethnicity, and age' (Andrews and Bonta 2010b, 47).

The categorization of individuals according to their risk level is based on predictive risk tools, of which the Level of Service Inventory-Revised (LSI-R), also developed by Andrews and Bonta, is the most studied and probably the most influential. The LSI-R is used by probation agencies in Canada, the U.S., the U.K., several European countries, Hong Kong and Australia (Hannah-Moffat 2013; van Wingerden et al. 2014; Monahan

and Skeem 2016). The tool has inspired OASys in England and Wales (Robinson 2017), which in turn has inspired the Dutch tool RISc (Risico Inschattings Schalen) and the Finnish tool RITA (Riski- ja tarvearvio) (van Wingerden et al. 2014; Salo et al. 2015). LSI-R is a 'risk/needs assessment tool' that predicts general recidivism. Because Andrews and Bonta were not only interested in risk but in opportunities for rehabilitation, they included 'dynamic' risk factors – also called 'criminogenic needs' – that are changeable and correlated to recidivism. Therefore LSI-R is considered a 'third-generation' tool (for detailed overviews of the evolution of risk assessment tools, see Hannah-Moffat 2005; Harcourt 2008). The LSI-R aims to measure the major risk and needs factors, the 'Central Eight', for both assessing risk of recidivism and targeting criminogenic needs (Andrews and Bonta 2010a). The 'Central Eight' consist of the 'Big Four' factors (History of Antisocial behaviour, Antisocial Personality Pattern, Antisocial Cognition and Antisocial Associates) and the 'Moderate Four' factors (Family/Marital Circumstances, School/Work, Leisure/Recreation and Substance Abuse). According to Andrews and Bonta, the Central Eight are the best predictors to future offending and are criminogenic needs that can be altered through rehabilitation in order to decrease reoffending.

A major concern in the debate about predictive justice is bias based on social marginality, race/ethnic background, and gender (e.g. Harcourt 2008; Starr 2014; Hannah-Moffat 2016; van Eijk 2017). As long as predictive tools exist, social marginality has played a role in predicting risk, although some risk tools have eventually excluded socioeconomic factors due to their correlation to race (Harcourt 2008; Tonry 2019a). In current general risk assessment tools, such as the LSI-R but also COMPAS, OASys and RISc, socioeconomic bias is 'built-in' because they measure socioeconomic items to calculate risk scores (van Eijk 2017). In addition to items that measure socioeconomic status directly, other items correlate with socioeconomic marginality, such as leisure activities (depend on money, time and geographical location) and attitudes (trust in police is affected by actual experiences with police, and relations with police tend to be problematic in poor neighbourhoods and communities of colour) (Harcourt 2008; Goddard and Myers 2017). Furthermore, qualitative evaluations of attitudes and lifestyle require human interpretation, which make standardized assessments vulnerable to subjective judgment and thus class prejudice or implicit bias (Hannah Moffat et al. 2009).

While it may seem evident to many scholars, policy makers and practitioners that socioeconomic factors are criminogenic, there is reason to reconsider. Andrews and Bonta (2010a) acknowledge that school and work are among the 'Moderate Four' factors of the 'Central Eight' criminogenic factors, whereas Taxman and colleagues (2014) consider school, work, housing and neighbourhood as non-criminogenic factors that indirectly impact offending behaviour. Monahan and Skeem (2016, 19) similarly argue that understanding criminogenic needs as causal factors is highly questionable, as we lack rigorous experimental studies to determine causality (and it many cases it would be unethical to design randomized controlled trials). Yet, including socioeconomic factors in predictive algorithms is a 'fundamental conceptual layer of bias' (Eckhouse et al. 2019) which by and large is accepted by many practitioners and experts. This is particularly problematic given that socially marginalized individuals are overrepresented in Western criminal justice systems, even in relatively equal societies such as the Netherlands. Predictive justice thus affects marginalized groups most and it affects them most negatively, as decisions are made based on social marginality. Socioeconomic disparities, and consequently racial/ethnic and gender disparities in criminal-legal decisions may thus

be exacerbated by predictive algorithms (van Eijk 2017). This has not led to abandoning socioeconomic items from risk tools.

In the next sections I investigate how the human-algorithm interplay produces the 'high risk' criminal subject in algorithmic practices of rehabilitation in more detail, paying specific attention to the role of social marginality.

The production of the abstracted 'high-risk' subject

Rehabilitation according to the RNR model starts with predicting an individual's risk of reoffending based using an algorithmic tool. Algorithmic prediction is a process that simultaneously deindividualizes and individualizes risk. It deindividualizes risk because actuarial assessment relies on aggregate data and averages (Hannah-Moffat 2013; Miller and Morris 1988). While concrete decisions about individuals are 'individualized', risk assessment is not: 'the riskiness attributed to an individual is not his or her own, but the average of a group in which he or she is included for purposes of statistical analyses' (Tonry 2019a, 446). The abstracted risk score should be distinguished from 'individual risk intrinsic to the subject herself' (ibid.). Punishing individuals based on aggregate statistics has been criticized on theoretical, methodological and ethical grounds (e.g. Hannah-Moffat 2005, 2013; Harcourt 2008; Simmons 2018). It may still be useful for decision-makers to know the reoffending rate of categories of individuals that are similar to a particular individual (Monahan and Skeem 2016), but in essence it is impossible to say anything meaningful about the future behaviour of one individual.

However, from studies on how practitioners interpret algorithmic predictions we know that professionals tend to interpret risk scores in such a way that risk becomes intrinsic to the individual. Hannah-Moffat (2013, 278) observes that 'despite being trained in the use and interpretation of risk tools, practitioners tended to struggle with the meaning of probability scores':

> 'Instead of understanding that an individual with a high risk score *shares characteristics* with an aggregate group of high-risk offenders, practitioners are likely to perceive the individual *as* a high-risk offender. In practical terms, *correlation becomes causation* and potential risk is translated into administrative certainty (ibid., italics in original)'.

Thus, in interpreting algorithmic outcomes, causal explanations are inferred from correlations and because an individual is assessed as 'high risk' they are assumed to be more dangerous to society (ibid.; see also Harcourt 2008). In his study on parole in California, Werth (2019) similarly observes that the abstracted risk assessment becomes an individualized evaluation: the individual that is assessed as 'high risk' becomes an inherently dangerous subject (Werth 2019). In line with these observations, Monahan and Skeem (2016) describe a tendency among judges to conflate risk and blame, while these judgements should play a role in different types of decisions: blame is relevant for conviction while risk is relevant only for sentencing.

These interpretative slippages are not solely the product of the use of algorithms; they are inherent to risk thinking. According to Werth (2017, 822), risk assessment is the result of a 'techno-moral assemblage': practitioners bridge and integrate different 'ways of knowing (file-based, actuarial, experiential, moral and affective)' (see also Hannah-Moffat et al. 2009). However, even if the actuarial way of knowing is one among other

ways, the use of algorithms seems to impact other ways of knowing. Such technologies are seen as rational, scientific and value-neutral and therefore fair and just – and for these reasons fairer and more just than humans (Silver 2000). This 'rationalization process' works to conceal the ways in which criminal justice policies and practices engage in social classification that distinguishes deserving and undeserving offender categories (ibid.; Lamont et al. 2014). Rationalization also tends to push aside ethical questions about procedural and individual justice and social consequences such as the reproduction of inequality.

Silver (2000) has pointed out that the social sciences contribute to rationalization in at least two ways: first, by advancing risk assessment technologies for managing populations, and second, by providing an interpretive framework which justifies population management. The RNR literature is a case in point that 'the causes of crime claimed by theory and research underpin and provide a framework for actuarial assessments or risk-based technologies' (Metcalf and Stenson 2004, 8). Indeed, manuals for algorithmic tools present them as insightful for understanding not just correlation but causation. For example, the OASys manual published by the British National Offender Management Service states that OASys can be used to 'help assessors in understanding the ''why'' of offending' (NOMS 2009, 8) and the COMPAS manual details many criminological theories, from psychological to sociological and situational theories, that 'help us understand more about why people make their behavioural choices' (Northpointe 2012, 6). Given this contradictory information and lack of statistical expertise, it is not surprising that there is tendency among judges, parole and probation officers to understand risk factors as causal factors that explain an individual's behaviour.

The focus on risk thus has not displaced a focus on the causes of crime. Rather, risk factors are interpreted as causal to offending behaviour. Like risk scores, risk factors should be understood as 'statistical predictions', not actual predictions about an individual, let alone as causal factors. But in the context of rehabilitation it seems virtually impossible not to draw the erroneous conclusion that probabilities are certainties and that risk factors are causal to offending, as insights into risk scores and risk factors serve the purpose of informing correctional treatment (Hannah-Moffat 2013). Even more than in the context of sentencing, for decision-makers involved in rehabilitation the individual with a high risk score is likely to become *known* as a high-risk offender based on identified criminogenic characteristics and problems.

Understanding risk factors as explanations for behaviour is particularly problematic when these factors are part of an individual's identity or lie outside an individual's scope of action. Starr (2014) problematizes including factors such as age and gender, but also employment status and residential neighbourhood, for it makes punishment dependent upon a person's identity rather than what the person has done. For Tonry (2019b) this is one of the problems that leads him to conclude that predictive sentencing is morally unjustifiable, although his argument is slightly different. Tonry considers it 'per se unjust' to use variables such as age, gender and race in predictive sentencing because individuals have no control over them or at not morally responsible for them, but for socioeconomic factors, Tonry argues that:

> 'These are quintessentially *personal choices*; people in free societies are entitled to make those decisions for themselves and not to suffer because of the choices they make (2019b, 14-15, italics added).

Within the RNR model, Andrew and Bonta's (2010a) argue for including 'socioeconomic achievement' in risk/needs assessment because it would be an 'achieved' status, as opposed to social class (family background) as an 'ascribed' status. It is, however, questionable that current socioeconomic circumstances are 'quintessentially personal choices' or that they are achieved rather than ascribed. Decoupling socioeconomic achievement from one's social class ignores that even in relatively equal and egalitarian societies intergenerational social mobility is more limited than the meritocratic ideal presents it (e.g. Corak 2013). And while it is true that individuals cannot be punished for choosing unemployment, as Tonry argues, this argument ignores the structural factors that impact an individual's educational level, their socioeconomic opportunities, housing situation or residential neighbourhood: the role of parental socioeconomic status, limited social mobility, austerity, spatial segregation, employment rates, the rising costs of housing and living expenses, and unequal access to education.

Here we see the process of how risk factors are simultaneously deindividualized and individualized. Predictive justice produces an understanding of social marginality as 'decontextualized': it disregards that social marginality is shaped by socio-structural inequality (Hannah-Moffat 2016; Goddard and Myers 2017). Through decontextualization, the 'high-risk' subject is made individually responsible for their risk factors, which conveys the message that individuals are to blame not only for their offending but also for their socioeconomic marginality (van Eijk 2017). Put differently, risk assessment produces the individual as an inherently 'high risk' subject, responsible for their own marginalized position, ignoring the context in which education, employment as well as attitudes and coping mechanisms – criminogenic or protective factors – are shaped. Andrews and Bonta (with Wormith 2011) have responded to critics that their RNR model does consider contextual factors because it addresses employment, education and such. However, it is not sufficient to account for these circumstances as contextual to behaviour – these circumstances should be reviewed in their wider socio-political context as well, to account for the limited influence of individuals over their own circumstances.

Addressing racial inequality, but we may consider this more broadly, Hannah-Moffat (2013, 282) suggests that 'it is impossible to treat individuals fairly if they are treated as abstractions, unshaped by the particular contexts of social life'. The problem of de/individualization of risk and risk factors is inherent to what algorithms aim to do – predict future behaviour – and to how algorithms work – analysing aggregate, not individual, data. In addition, humans have trouble understanding how probabilities relate to individuals. Their understanding of risk and risk factors as inherent to an individual violates with the notion of individual justice: individuals are not really evaluated on their own merits. In the next section I investigate how this may impact human dignity and the process of desistance.

Re-envisioning a 'low-risk' future self *against all odds*?

After an individual is assessed and labelled as 'high risk' subject who is inherently dangerous due to risk factors which are seen as causal to their behaviour, individuals are expected to engage with the process of rehabilitation and ultimately to desist from offending behaviour. While we know very little about how justice-involved individuals experience algorithmic evaluation, studies on desistance provide relevant insights into how individuals may negotiate algorithmic predictions about their future. Researchers

agree that desistance follows from a combination of structural and subjective factors, although theories may emphasize one or the other as the most important catalyst for change (Bersani and Doherty 2018; Farrall et al. 2011). In any case, it is clear that desistance requires 'agency', a sense of having some degree of control over the future direction of one's life (King 2012; Healy 2013; Farrall et al. 2011). Agency is future oriented, it involves reflection on the past but also hope and optimism, a readiness for change (Healy 2013). Individuals may re-envision their past self and, more or less intentionally, construct a new self that is different, leaving behind their past self (Bersani and Doherty 2018; Hunter and Farrall 2018; McNeill 2006). According to Paternoster and Bushway (2009), in the process of desistance individuals imagine different possible selves and must in the long run find a balance between positive and negative possible selves. It is important to understand that re-envisioning a 'future self' requires effort and time and that there may be a tension between the current self and the future self which can cause frustration (Hunter and Farrall 2018; Harris 2011).

However, while the process of desistance requires re-envisioning a future self that is different from the past (or current) self, predictive algorithms predict a future self that is at high risk to reoffend and thus to stay the same. This dilemma is illustrated by an anonymous writer to *InsideTime* who responds to Hanson's letter about the dehumanizing effects of risk assessment and proposes to put OASys to the test:

> 'It is time to elect a new prime minister/world leader and only your (OASys) vote counts.
> Candidate A associates with crooked politicians and consults with astrologists; has had two mistresses; he also chain-smokes and drinks 8–10 Martinis a day.
> Candidate B was kicked out of office twice; sleeps until noon; used opium in college and drinks a quart of whiskey every evening.
> Candidate C is a decorated war hero; he's a vegetarian; doesn't smoke; drinks an occasional beer and has never committed adultery.
>
> Which candidate would be chosen by "OASys scoring"?
>
> Candidate A is Franklin Roosevelt; Candidate B is Winston Churchill; Candidate C is Adolf Hitler. (…) Does OASys work?' (Anonymous 2010)

The writer's thought experiment could be read as a critique on the poor predictive accuracy of algorithmic tools, but it also echoes Hanson's concern that predictive justice 'distorts what makes up the human condition and the potential for change'. People can and do change, but can we re-envision radically different futures for these three candidates based on their past behaviour? If algorithmic tools predict that an individual will likely be a 'persister' in the future, how does this shape how they – and others – imagine a different future self as a 'desister'? In other words, is algorithmic justice compatible with the principle of human dignity (cf. Ward 2011)?

The role of algorithmic prediction of risk may have a significant impact on re-envisioning potential future selves. Paternoster and Bushway (2009) argue that an individual's 'feared self' supports the initial motivation to change, as an individual does not want to become what they fear they will become. In this way, the production of the 'high risk' subject through algorithmic reasoning could stimulate agency and desistance. However, other scholars warn of the limiting force of assessing future risk of reoffending, as 'risk' may keep individuals trapped in past behaviours. McNeill and colleagues (2014), for

example, are critical of how the RNR rehabilitation model and its focus on individual risk factors impacts desistance as it is retrospective, whereas desistance requires a prospective outlook. Risk assessment informs the imagined possible future selves as it calculates the probability of a certain future in which the individual repeats similar behaviour in the future, based on past behaviour. This may be counter-productive to desistance, McNeill (2006, 53) argues, as measures that label, exclude and segregate 'seem designed to confirm and cement "condemnation scripts" and thus to frustrate desistance.' Risk assessment may limit the imagination of a positive future self by assuming that past behaviour continues in the future, especially when risk factors – erroneously interpreted as causal to the undesired behaviour – are difficult to change.

Furthermore, the presentation of algorithmic tools as scientific, evidence-based and thus accurate – even though predictions are never fully accurate – may strengthen the sense that the 'feared future self' is unavoidable. For a 'feared future self' to motivate individuals to change, there should be space for people to imagine that they will not become their feared future self. Put differently, a negative possible self is unlikely to motivate change of it means that individuals must re-envision change *against all odds*. Desistance is difficult as is, and it can become even more difficult if it involves 'proving the computer wrong'. Whether individuals are discouraged or encouraged by future images of the self as likely reoffender depends on how risk predictions are interpreted by both the practitioner and the individual subjected to algorithmic prediction. To respect human dignity, acknowledge potential and foster change, it is essential that algorithmic predictions are taken as probabilities, not certainties. Certainties cannot be changed, while probabilities offer alternatives and thus opportunities for change. It is thus essential that rehabilitation practitioners avoid the interpretative mistakes that the studies described earlier have observed: to respect human dignity, they should resist the tendency to individualize the patterns that are found on an aggregate level.

Desistance studies furthermore show how structural factors may hinder and discourage individuals from reconstructing their identity. Desistance is according to Farrall et al. (2011, 224) 'best approached as the result of the interplay between individual choices and a range of wider social forces, institutional and societal practices which are beyond the control of the individual.' In other words, agency is enabled or constrained by structural factors and social context. Maruna (as quoted in McNeill 2006) describes how both persisters and desisters have a sense of 'fatalism' in how they see their criminal career – as an individual overcome by criminogenic structural pressures. Desisters succeed in '"discovering" agency in order to resist and overcome' these pressures, while persisters viewed their future as a script that has been written for them long ago (ibid., 48). The process of changing one's identity and one's behaviour may thus feel as if individuals are 'changing fate' (Healy 2013). Desistance requires optimism about one's future, perhaps more optimism than a risk score would warrant (Cobbina and Bender 2012). Harris (2011) observes that the motivation of individuals to change can be short-lived, as they are confronted with structural forces as barriers to change – for example, limited options for housing, work, family support and the stigma of a criminal record. As 'individuals could go only so far in creating these structural changes' (ibid., 83), practical support is needed to keep individuals motivated.

It is possible that individuals may come to experience algorithmic predictions as structural (institutional) forces that hamper their imagination and ability to change. If in everyday situations individuals feel that algorithms reduce them to a number (Binns et

al. 2018) and have negative emotional responses to algorithmic decisions (Lee 2018), it is not difficult to imagine that individuals in the criminal justice context feel negatively about algorithmic justice as well. Prediction scores add yet another factor over which individuals have limited control – especially when decisions are black-boxed and practitioners only provide insight into the risk score and priorities co-produced by the tool (Hannah-Moffat et al. 2009). Defending the RNR model, Andrews, Bonta and Wormith (2011, 742) argue that in order to motivate individuals, practitioners should

> 'Offer the client important information regarding the findings of the assessment. Open reviews of the results and implications of RNR-based assessments with moderate and higher risk individuals are a fundamental approach to the initiation of mutually agreed on service plans.'

However, discussing outcomes with individuals but not the way in which prediction works, may result in viewing predictions as fixed rather than seeing them as probabilities.

Especially for marginalized offender categories, algorithm predictions may be discouraging, as their risk score depends in part on their marginalized position in society. Education, employment and housing are at least partly outside an individual's scope of action and thus difficult to change in a way that supports identity change. Yet, through algorithmic reasoning marginalized individuals are made responsible for their criminogenic factors and expected to imagine a positive possible self that can negotiate structural inequalities (Hannah-Moffat 2016). Reviewing the potential of criminogenic needs for treatment, Taxman et al. (2014) conclude that socioeconomic factors should be considered only as 'stabilizers' to an individual's life. Given structural social inequalities, it seems fairer to consider socioeconomic factors as non-criminogenic factors that are contextual and not causal to desistance. Such an approach to socioeconomic marginality as criminogenic need may prevent that algorithmic predictions become experienced as unavoidable futures that discourage individuals to re-envision a different self and future. In this context we should also consider that in many jurisdictions individuals are subjected to multiple predictive tools during their involvement in the criminal justice system, from pretrial detention to release (van Eijk 2020). Prior predictions that have informed decisions on freedom and resources may produce criminogenic circumstances that, once an individual enters treatment, may work as structural obstacles to agency and desistance (Hannah-Moffat 2016). In short, studies on the process of desistance demonstrate that change is difficult for structural and subjective reasons, and we need to investigate the real possibility that involving algorithms makes this process even more difficult for individuals.

Principles for algorithmic justice

It is clear that algorithmic practices have transformed practices of justice, although many questions remain about how algorithms are used in practice and how they impact principles of justice and fairness. Based on what we know about risk-based justice in rehabilitation, I have in this Chapter argued that what algorithmic justice aims to do – predict future behaviour – and how it does so – based on aggregate, group data – contradicts notions of individual justice (that individuals should be evaluated based on their own behaviour) and, consequently, human dignity (that individuals have agency: a sense of control over their own future). However, the most fundamental problems of

algorithmic reasoning are the result of human decisions and human interpretation rather than a problem of technology. Ward (2011) argues that predictive justice requires 'extreme care' in gathering data about criminogenic needs and 'that any subsequent decisions that restrict offenders' freedom, and possibly well-being, are rationally (and ethically) justified'. While in principle I agree, I would also caution against focusing on rational justifications, as they tend to push aside ethical considerations. The RNR model and measures for the accuracy and effectiveness of predictive tools offer rational justifications that invite us to treat algorithmic justice as evidence-based and data-driven and thus authoritative, value-free and a-political evaluations of behaviour and its causes. In practice, however, algorithms are imbued with value, as decisions about how to use algorithms 'are, in the end, not a statistical or scientific matter, but a political and social judgment about what risks are unacceptable, and what responses to risks should be allowed' (Miller and Morris 1988, 268). The question whether prediction should play a role at all in criminal-legal decisions is fundamentally an ethical question but is usually treated as a technical issue of accuracy. Categories of 'low', 'medium' and 'high' risk are constructed and for each category 'false positives' and 'false negatives' are accepted as unavoidable. Similarly, including certain risk factors or not, and which arguments are put forward by academics, experts and decision-makers, is a moral-political negotiation – for example, including racial factors in risk assessment is now seen as unacceptable, while socioeconomic factors are still seen as acceptable despite concerns of class, racial/ethnic and gender bias. Rationalization seems to neutralize concerns about constructing social marginality as risk factor and thus facilitates the alignment of technological approaches with political interests (cf. Winner 1980), as social marginality is decontextualized and individualized through algorithmic reasoning, making individuals fully responsible for their own (deprived) situation.

It is essential to understand the human-algorithm interplay within the context of specific practices, as predictive sentencing presents different problems than predictive rehabilitation. To investigate and improve the ethics of algorithmic justice, we should thus consider algorithmic technologies not as independent from human practices but as integrated with it and view algorithms as reflections and extensions of human practices. Many fundamental problems of predictive justice arise from the problematic nature of risk-thinking and from erroneous interpretations of algorithmic prediction. It is crucial that practitioners are trained to interpret and explain algorithmic predictions, as experiences of individuals subjected to algorithms are mediated through a human expert who interprets algorithmic outcomes Furthermore, if practitioners make interpretative mistakes, it is likely that justice-involved individuals make similar mistakes. As decisions-makers in the criminal justice system are not experts on statistics, education could help to correct errors in interpreting algorithmic predictions or algorithms could be designed differently to improve interpretability (Green and Chen 2019). From the perspective of individuals who are subjected to algorithmic justice, more effort to explain not just outcomes but the logic of algorithmic prediction helps them to evaluate the fairness of algorithmic decisions (ibid.; Binns et al. 2018). This implies that the use of black-boxed algorithms – specifically algorithms that are secret due to their proprietary nature – is unethical. In the context of rehabilitation, practitioners have an ethical duty to explain algorithmic tools correctly to justice-involved individuals in order to avoid discouraging the re-envisioning a positive future self. Working towards an ethical use of algorithms requires addressing the inherent tensions between prediction on the one hand and individual justice and human dignity on the other. Rethinking this tension may result in a

decision to not use algorithmic justice in a rehabilitation context, as to fully respect human dignity and foster agency and personal change.

While there are many reasons to be concerned about algorithmic justice – not only in the context of sentencing but also in relation to rehabilitation – it has also created opportunities for critical evaluation by a wider group of experts of decisions in criminal justice that were long hidden from the public and considered the exclusive domain of legal professionals (Peeters and Schuilenburg 2018). The use of more advanced technology, embedded in broader concerns about digitalization and big data, have attracted new experts – statisticians, AI experts, investigative journalists, among others, in addition to social scientists and legal scholars – to scrutinize the design, use and consequences of predictive justice. This critical debate is mirrored by a growing academic and professional literature and unprecedented private sector involvement that reinforce the image of algorithmic justice as scientific and fundamentally unproblematic. The voices that are still missing from this debate are those of the individuals who are subjected to algorithmic justice. If the act of prediction itself is perceived as unfair, no algorithm is going to make prediction fair. To further the debate about ethical use of algorithmic justice we need to include the experiences of individuals who are subjected to algorithmic tools in order to ensure that they experience algorithmic justice as fair and constructive to their rehabilitation and well-being.

References

Andrews D.A., and Bonta J. 2010a. *The Psychology of Criminal Conduct*. London: Routledge.

Andrews, D.A., and Bonta J. 2010b. "Rehabilitating criminal justice policy and practice." *Psychology, Public Policy, and Law* 16 (1): 39.

Andrews, D.A., Bonta, J., and Wormith, J.S. 2011. "The risk-need-responsivity (RNR) model: Does adding the good lives model contribute to effective crime prevention?" *Criminal Justice and Behavior* 38 (7): 735-755.

Anonymous. 2010. "'Standardised structured approach'". *InsideTime*, 1 June, https://insidetime.org/standardised-structured-approach/

Bersani, B.E., and Doherty, E.E. 2018. Desistance from offending in the twenty-first century. *Annual Review of Criminology* 1: 311-334.

Binns, R., M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (April) 377: 1-14. doi: 10.1145/3173574.3173951

Binns, R., 2019. "Human Judgement in Algorithmic Loops: Individual Justice and Automated Decision-Making." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3452030.

Bussert, T. 2019. "What the FIRST STEP Act Means for Federal Prisoners." *The Champion* (May): 28-36. Accessed 18 July 2019. https://www.frostbussert.com/files/what_the_first_step_act_means_for_federal_prisoners.pdf

Cobbina, J.E., and K.A. Bender. 2012. "Predicting the Future: Incarcerated Women's Views of Reentry Success." *Journal of Offender Rehabilitation* 51 (5): 275-294.

Corak, M. 2013. "Income inequality, equality of opportunity, and intergenerational mobility." *The Journal of Economic Perspectives* 27 (3): 79–102.

Desmarais, S.L., and E.M. Lowder. 2019. "Pre-trial risk assessment tools: A primer for judges, prosecutors, and defense attorneys." MacArthur Foundation Safety and Justice Challenge.

Eckhouse, L., K. Lum, C. Conti-Cook, and J. Ciccolini. 2019. "Layers of bias: A unified approach for understanding problems with risk assessment." *Criminal Justice and Behavior* 46 (2): 185-209.

Farrall, S., G. Sharpe, B. Hunter, and A. Calverley. 2011. "Theorizing structural and individual-level processes in desistance and persistence: Outlining an integrated perspective." *Australian & New Zealand Journal of Criminology* 44 (2): 218-234.

Goddard, T., and R.R. Myers. 2017. "Against evidence-based oppression: Marginalized youth and the politics of risk-based assessment and intervention." *Theoretical Criminology* 21 (2): 151-167.

Green, B., and Y. Chen. 2019. "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments." *Proceedings of the Conference on Fairness, Accountability, and Transparency* (January): 90-99. doi: 10.1145/3287560.

Hamilton, M. 2015. "Adventures in risk: predicting violent and sexual recidivism in sentencing law." *Arizona State Law Journal* 47 (1): 1-62.

Hannah-Moffat, K. 2005. "Criminogenic needs and the transformative risk subject: Hybridizations of risk/need in penality." *Punishment & society* 7 (1): 29-51.

Hannah-Moffat, K., 2013. "Actuarial sentencing: An "unsettled" proposition." *Justice Quarterly* 30 (2): 270-296.

Hannah-Moffat, K. (2016). "A conceptual kaleidoscope: Contemplating 'dynamic structural risk'and an uncoupling of risk from need." *Psychology, Crime & Law* 22 (1-2): 33-46.

Hannah-Moffat, K., P. Maurutto, and S. Turnbull. 2009. "Negotiated risk: Actuarial illusions and discretion in probation." *Canadian Journal of Law & Society* 24 (3): 391-409.

Hanson, C. 2009. "The dehumanising process of risk assessment". *InsideTime*, 1 December, https://insidetime.org/dehumanising-process-of-risk-assessment/

Harcourt, B.E. 2008. *Against prediction.* Online version. Chicago: University of Chicago Press.

Harris, A. 2011. "Constructing clean dreams: Accounts, future selves, and social and structural support as desistance work." *Symbolic Interaction* 34 (1): 63-85.

Healy, D. 2013. "Changing fate? Agency and the desistance process." *Theoretical Criminology* 17 (4): 557-574.

Hunter, B., and S. Farrall. 2017. "Emotions, future selves and the process of desistance." *The British Journal of Criminology* 58 (2): 291-308.

King, S. (2013). Transformative agency and desistance from crime. *Criminology & Criminal Justice* 13 (3): 317-335.

La Diega, G.N. 2018. "Against the Dehumanisation of Decision-Making." *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 9 (1): 3-34.

Lamont M., S. Beljean, and M. Clair. 2014. "What is missing? Cultural processes and causal pathways to inequality." *Socio-Economic Review* 12 (3): 573–608.

Lee, M.K. 2018. "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management." *Big Data & Society* 5 (1): 1-16.

Matzner, T. 2017. "Opening Black Boxes Is Not Enough–Data-based Surveillance In Discipline and Punish And Today." *Foucault Studies* 23 (August): 27-45.

McNeill, F. 2006. "A desistance paradigm for offender management." *Criminology & Criminal Justice* 6 (1): 39-62.

McNeill F., S. Farrall, C. Lightowler, and S. Maruna. 2014. "Desistance and Supervision." In *Encyclopedia of Criminology and Criminal Justice*, edited by G. Bruinsma and D. Weisburd, 958-967. New York: Springer.

Metcalf, C., and K. Stenson. 2004. "Managing risk and the causes of crime." *Criminal Justice Matters* 55 (1): 8-42.

Miller, M., and N. Morris. 1988. "Predictions of dangerousness: An argument for limited use." *Violence and victims* 3 (4): 263-83.

Monahan J., and J. Skeem. 2016. "Risk assessment in criminal sentencing." *Annual Review of Clinical Psychology* 12, 489-513.

NOMS. 2009. *Public Protection Manual. Chapter 9. Risk of Harm*. National Offender Management Service/HM Prison Service. 2020https://www.gov.uk/government/publications/public-protection-manual-chapter-9-risk-of-harm.

Northpointe. 2012. *Practitioners Guide to COMPAS*. Northpointe. http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf.

Paternoster, R., and S. Bushway. 2009. "Desistance and the 'feared self': Toward an identity theory of criminal desistance." *Journal of Criminal Law and Criminology* 99 (4): 1103–1156.

Peeters, R., and M. Schuilenburg. 2018. "Machine justice: Governing security through the bureaucracy of algorithms." *Information Polity* 23 (3): 267-280.

Robinson, G. 2017. "Stand-down and deliver: Pre-sentence reports, quality and the new culture of speed." *Probation Journal* 64 (4): 337-353.

Salo, B., T. Laaksonen, and P. Santtila. 2016. "Construct validity and internal reliability of the Finnish risk and needs assessment form." *Journal of Scandinavian Studies in Criminology and Crime Prevention* 17 (1): 86-107.

Silver, E. 2000. "Actuarial risk assessment: Reflections on an emerging social-scientific tool." *Critical Criminology* 9 (1–2): 123–143.

Simmons, R. 2018. "Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System." *U.C. Davis Law Review* 52 (2): 1067-1118.

Starr, S.B. 2014. "Evidence-based sentencing and the scientific rationalization of discrimination." *Stanford Law Review* 66: 803-872.

Stevenson, M.T., and J.L. Doleac. 2018. "*The roadblock to reform*." American Constitution Society. https://www.acslaw.org/wp-content/uploads/2018/11/RoadblockToReformReport.pdf.

Taxman F.S., M. Caudy, and S. Maass. 2014. "Actualizing Risk-Need-Responsivity." In *Encyclopedia of Criminology and Criminal Justice*, edited by G. Bruinsma and D. Weisburd, 1-11. New York: Springer.

Tonry M. 2019a. "Predictions of Dangerousness in Sentencing: Déjà Vu All Over Again." *Crime and Justice* 48 (1): 439-482.

Tonry, M. 2019b. "Fifty Years of American Sentencing Reform: Nine Lessons." *Crime and Justice* 48: 1-34.

van Eijk, G. 2017. "Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality." *Punishment & Society* 19 (4): 463-481.

van Eijk, G. 2020. "Inclusion and exclusion through risk-based justice: analysing combinations of risk assessment from pretrial detention to release." *British Journal of Criminology* (advance online publication). doi: 10.1093/bjc/azaa012.

van Wingerden, S., J. van Wilsem, and M. Moerings. 2014. "Pre-sentence reports and punishment: A quasi-experiment assessing the effects of risk-based pre-sentence reports on sentencing." *European Journal of Criminology* 11 (6): 723-744.

Ward, T. (2011). Human rights and dignity in offender rehabilitation. *Journal of Forensic Psychology Practice*, 11 (2-3): 103-123.

Werth, R. 2017. "Individualizing risk: Moral judgement, professional knowledge and affect in parole evaluations." *British Journal of Criminology* 57 (4): 808-827.

Werth, R. 2019. "Theorizing the performative effects of penal risk technologies: (Re)producing the subject who must be dangerous." *Social & Legal Studies 28* (3): 327-348.

Winner, L. 1980. "Do artifacts have politics?" *Daedalus* 109 (1): 121-136.