# The Future of Mental Health, Disability and Criminal Law

Essays in Honour of Emeritus Professor Bernadette McSherry

**Edited by**
**Kay Wilson, Yvette Maker, Piers Gooding and Jamie Walvisch**

First published 2024

ISBN: 978-1-032-39607-1 (hbk)
ISBN: 978-1-032-39632-3 (pbk)
ISBN: 978-1-003-35064-4 (ebk)

## 13 The Digital Turn in Mental Health and Disability Law

### Actuarial Traditions and AI Futures of Risk Assessment From a Human Rights Perspective

*Piers Gooding and Yvette Maker*



DOI: 10.4324/9781003350644-17

The funder of the Open Access of this chapter is Australian Research Council

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

# 13 The Digital Turn in Mental Health and Disability Law

## Actuarial Traditions and AI Futures of Risk Assessment From a Human Rights Perspective

*Piers Gooding and Yvette Maker*

### Introduction

In 2019, senior members of the Trump administration, including the then President Donald Trump and Vice President Mike Pence, were briefed on a proposal called, 'SAFEHOME – Stopping Aberrant Fatal Events by Helping Overcome Mental Extremes' (Wan 2019). SAFEHOME involved experimentation to explore whether 'technology including phones and smartwatches can be used to detect when mentally ill people are about to turn violent' (Wan 2019). The programme would sit within a proposed new research arm of the US government called the 'Health Advanced Research Projects Agency', to be modelled after the Defence Advanced Research Projects Agency (DARPA).

SAFEHOME was not adopted during Donald Trump's term in office. Yet, the proposal highlights the proximity of ideas about algorithmic risk management with policies concerning mental health, 'preventive policing' and emerging technologies of biometric monitoring and surveillance. SAFEHOME builds on experiments in preventive policing in recent years, particularly in the US, which involve flows of mental health-related data from disparate public and private systems. In 2018, for example, the Florida state legislature authorised the collection and digitisation of certain types of student mental health data and its distribution through a state-wide police database (Travis 2019). Authorities reportedly intended to correlate the data with social media monitoring activity, with the purported aim of preventing gun violence (Travis 2019). Also in 2018, municipal police in Canada collated non-criminal information about individuals who had self-harmed or attempted suicide (Office of the Privacy Commissioner of Canada 2017). The information was then circulated to US border authorities, who used it to deny several Canadians entry into the US (Office of the Privacy Commissioner Canada 2017). These examples, too, highlight the growing potential for risk management in mental health policy and in policing more generally to be linked to the automated and data-driven technologies that power the contemporary information economy.

Our mentor and colleague Bernadette McSherry has explored such links, drawing on her longstanding work on predictive analytics in assessment tools

DOI: 10.4324/9781003350644-17

that are used both to define and diagnose mental health conditions and in efforts to predict future violence or recidivism among individuals deemed to pose a risk (see e.g. McSherry 2014, 2017; McSherry and Freckelton 2013; McSherry and Keyzer 2011). McSherry's more recent works on the technological frontiers of these developments have examined efforts to detect and treat mental health conditions using social media data and other health-related datasets (McSherry 2018) and the use of algorithms in forensic psychiatry to predict risk of harm (McSherry 2020). McSherry's work so far has been unique in applying a human rights lens to these technological developments in forensic mental health. In this chapter we seek to extend on that work and, in doing so, contribute to the growing, global conversation on the implications of data-driven, automated and AI-enabled technology in the mental health and disability contexts (see e.g. Gooding and Kariotis 2021; Gooding and Resnick 2020; Human Rights Council 2021; Marks 2019, 2020; Paterson and Maker 2021; Valentine, D'Alfonso and Lederman 2022), as well as in criminal law and health and social care more generally (see e.g. Eubanks 2018). Our aim is to extend the discussion to the range of techniques that use – or claim to use – automated or AI-enabled technology to assess mental health-related risk. Due to the 'black box' and proprietary nature of much of this technology, it is not always clear what processes are in use, but they range from the automation of statistical or actuarial processes to more sophisticated technologies that use machine learning, neural networks or natural language processing to make predictions, recommendations or decisions (OECD 2019). These techniques and devices share both similarities to, and differences from, earlier risk assessment techniques that utilised actuarial methods and algorithms to produce estimates of individuals' likelihood of reoffending or other forms of 'riskiness'.

We draw on McSherry's work to review technological developments in forensic mental health risk assessment and consider critiques of these developments. This includes concerns raised about the use of automated and AI-enabled technology in risk assessment that build on concerns about traditional forms of risk assessment, as well as critiques that draw on human rights principles and laws. We then connect this work with recent discussion about the contribution and limitations of a human rights-based approach, particularly compared to an approach that foregrounds ethics, to critically assessing the social and other human consequences of AI and automated decision-making. Arguments in favour of such an approach are growing in volume, largely on the basis that human rights provide widely accepted normative standards against which to assess these consequences and mechanisms for overseeing their use, which more conventional analyses based on principles of 'ethical' AI arguably do not. Other scholars and commentators seek alternative political, legal and ethical approaches that aim to move beyond the binary confines of ethics and human rights (see e.g. Benthall and Goldenfein 2020; McQuillan 2022), though we contain our focus for the purposes of this chapter to ethics and human rights as proposed governance or analytical frameworks.

We conclude that the use of automation and AI-enabled technology in risk assessment in forensic mental health contexts raises substantially the same issues that already exist in longer standing critiques of traditional risk assessments. These principally relate to bias, confounding variables and issues of transparency. However, we also consider that the extent or nature of some of these issues is significantly different in several senses. This includes differences associated with the speed and scale of decisions that are possible (such as automated risk assessment of arrested individuals awaiting trial based on massive, population-level datasets) the opacity of AI and other automated approaches that add to existing problems with transparency in risk assessment, and the major legal implications in areas such as privacy and discrimination that are raised by the monitoring and surveillance capabilities of new biometric technologies that may be used in risk assessment. We argue that efforts toward accountability in the use of these risk assessment technologies must address the needs and rights of people with disability, including mental health-related or psychosocial disability, in order to respond to growing calls for societal use of automation, AI and other data-driven technologies in such a way that protects and promotes the human rights of these groups on an equal basis with others (Whittaker et al. 2019; Human Rights Council 2021).

## Actuarial Risk Assessment in Mental Health and the Digital Turn

The use of predictive analytics in forensic mental health settings is not new. By 1990, according to Robert Castel (1991), risk management had become a core professional responsibility of all those involved with psychiatry. Various techniques were designed to identify and measure levels, signs and indicators of 'risk' or 'dangerousness'. These techniques built on a longer tradition in the criminal law context. As far back as the 1920s, statisticians in Europe and North America were taking steps to formalise the assessment of parolees' likelihood of recidivism (Mathiesen and Rutherford 2006: 95).

Early risk assessment techniques in forensic mental health and other criminal justice settings involved actuarial methods based on regression, and then evolved to relying on algorithmic risk assessment that provides a probabilistic estimate of the likelihood of reoffending (Stevenson and Doleac 2019; McSherry 2020). Slobogin (2012: 198) has explained that the latter involve 'actuarial prediction devices that rely on empirical discovery of factors associated with recidivism, which are then weighted and combined according to an algorithm that produces recidivism probability estimates'. Such predictive work has served several purposes, including informing bail or parole decisions and determining whether a person should be subject to 'preventive justice', such as post-release monitoring or post-sentence detention and supervision (McSherry 2020). Proponents offered such approaches as a way for courts and mental health professionals to improve on 'unstructured' decision-making. Unstructured decision-making in risk assessment, which essentially refers to

discretionary judgments by authorised individuals, is broadly considered to be 'notoriously bad, biased, and reflexive, and often relies on stereotypes and generalizations that ignore the goals of the system' (Slobogin 2021: ix). The aim of standardised risk assessment has been to replace such an approach with one that is efficient, predictable, accurate and consistent, and hence to aspire to 'objectivity' and 'certainty in decisions about preventive justice' (McSherry 2020: 30). It has also been suggested that some newer systems, such as those that use biometric monitoring techniques to make a human 'machine-readable', can circumvent the need for subjective patient reporting or clinical observation (see e.g. Resnick and Appelbaum 2019). This latter view carries controversial value judgments about the relative merit of 'overcoming' the subjective viewpoints of an individual and their mental health practitioners, in favour of 'objective' biometric data about that person, such as voice activity, bodily motion and heart and respiratory rates (Resnick and Appelbaum 2019; cf. McQuillan 2018). We will discuss such biometric or 'passive monitoring' shortly.

The first two decades of the 21st century have seen risk assessment for the purpose of managing individuals rise in prominence in mental health service provision in many countries (Holmes 2013; Slemon, Jenkins and Bungay 2017) and throughout society (Lupton 2013). Today, psychiatric services in many parts of the world are centrally concerned with pre-emptive interventions to prevent some future harm (Mossman 2009; Holmes 2013; Szmukler and Rose 2013; Slemon, Jenkins and Bungay 2017). In the US, for example, Christopher Slobogin (2021: vii) notes that almost every state has 'authorized the use of algorithms that purport to determine the recidivism risk posed by people who have been charged or convicted of crime'.

While it may not be new for government agencies, health providers and private companies to collect and hold large volumes of information about people, or to attempt to make individuals more 'calculable' (Miller and Rose 1995), the quantity and detail of the data which they are now able to collect – and the rapid speed with which it flows through complex communication ecosystems – is unprecedented. In other words, the preoccupation with risk among mental health practitioners has expanded alongside the acceleration of digital approaches to health care, making it unsurprising that steps have been taken to digitise predictive analytics in risk assessment and management (McSherry 2020).

Contemporary data-based approaches to risk assessment can deal in ever-expanding datasets, increasingly complex algorithms and exponential increases in computational power. These deterministic systems range from those that employ relatively simple binary logic, through to those that utilise 'deep learning' or machine learning which make probabilistic predictions based on complex algorithms and massive datasets (Zalnieriute, Bennett Moses and Williams 2019: 432). Some approaches just automate actuarial and statistical processes of the past. Others seek to identify new patterns or correlations in data and formulate rules on that basis. Experiments currently underway include the

application of natural language processing and machine learning to forensic eHealth records (Le et al. 2018) and the use of machine learning on patient registries using a range of sociodemographic, judicial and psychiatric variables to determine whether 'early intervention' should occur (Trinhammer et al. 2022).

As noted, experiments are underway to use automated software to classify human behaviours using remote technologies to monitor or surveil individuals. One iteration, 'passive monitoring', refers to generation of data about individuals that does not require them to actively respond (Resnick and Appelbaum 2019), as compared to 'active monitoring' where a respondent might provide mood self-reports. In one Austrian study, ten patients in a psychiatric hospital were given a mobile phone with an application (or app) that was 'based on smartphone behavior and activity monitoring' over a 12-month period (Grünerbl et al. 2015). The app generated data, used with the consent of patients, that was 'usable [by clinicians] as an "objective" measurement that help[ed] detect state changes to guarantee the availability of in-time treatment' (Grünerbl et al. 2015: 142). The researchers compared evaluations of patient mood using sensors that generated information concerning phone-use, voice and bodily movement, to evaluations using standard scales (such as the Hamilton Depression Scale). The results indicated a 'state change detection precision and recall of over 97%' (Grünerbl et al. 2015: 142), suggesting close to perfect accuracy.

In the context of the US, Kimberly Resnick and Paul Appelbaum (2019: 457) have explored the implications of being able to set 'objective parameters that correlate with mental health status and create an opportunity to use Big Data and machine learning to refine diagnosis and predict behavior'. They consider how these approaches could be applied to the forensic mental health and criminal law context. Examples include:

- accused persons who are found not guilty by reason of mental impairment being subject to continuous monitoring and real-time analysis of body data that generate 'electronic biomarkers' of relapse in ways that 'bypass the limitations of approaches that rely on self-report and clinical interview of insanity acquittees';
- defendants who have been diverted by the legal system into specialty mental health courts being subjected to passive monitoring arrangements, which again, monitor for relapse; and
- 'offenders with mental illnesses [more generally being] given the option of passive monitoring or entering or remaining in confinement' (Resnick and Appelbaum 2019: 461–63).

These possibilities highlight a complicating dimension to the discussion of forensic mental health risk assessment today; namely, that recent technological developments create the possibility of ongoing, passive monitoring and surveillance of persons through remote technologies and through real-time,

automated, algorithmic analysis of the data generated. In such circumstances, the assessment of some future harm is no longer conducted at a single point in time, such as during a pre-trial or post-conviction assessment by a psychologist or psychiatrist. Instead, remote risk assessment can occur on a continuous basis. This could be posited as a logical (and individualised) progression from current electronic monitoring of convicted persons and forensic mental health patients through the use of radio-frequency or GPS data (see Miller 2015; Bartels and Martinovic 2017).

Notably, continuous, remote monitoring or 'passive monitoring' operates in ways that go beyond assessment, having the potential to generate more than merely clinical data (notwithstanding that what counts as 'clinical data' may be contested). For example, Resnick and Appelbaum (2019: 463) point out that the generation of a body of data through compulsory monitoring would be almost 'irresistible to law enforcement and prosecutors', at least in the US, given that such data 'is likely to be considered nontestimonial by the courts and thus will likely fall outside the bounds of psychotherapist – patient privilege, even if it is being gathered for clinical purposes'. These possibilities start to break down the distinction between risk assessment (taking steps to evaluate the likelihood of future offending) and risk management (taking steps to reduce risk). For a related discussion, see Chapter 10 by Christopher Slobogin in this volume.

The use of passive monitoring in contemporary risk assessment, therefore, raises the much larger issue of biometric monitoring and surveillance, where 'biometric monitoring' refers to the tracking of individuals' characteristics related to changes in human traits and body parameters, such as bodily motion, location, voice activity, and heart and respiratory rates. There is a growing body of law reform and research activity on the legal, ethical, political and social dimensions of biometric monitoring and surveillance (see Kak 2021), including criticism of the epistemological aspiration to 'overcome' the subjective account of the individual or care professional in favour of purportedly objective technical measures (McQuillan 2018). For the purpose of this chapter, what is important is that both forms of risk assessment, whether single-point-in-time assessment or 'remote, continuous assessment' (and management) using biometric monitoring and surveillance, rely on statistical methods and probabilistic reasoning in predicting some future harm. These methods and forms of reasoning regarding risk assessment have been the subject of criticism from multiple angles, even as others have pointed to benefits, such as greater accuracy and the reduction of bias, that may be made possible by such methods.

## Critique of Risk Assessment in Mental Health Settings: Old and New

McSherry notes at least three criticisms of forensic mental health risk assessments that are especially pertinent to the more recent developments in

data-driven and AI-enabled technologies. First, she draws critical attention to the specific variables used in risk assessment, and the variable-based approach itself (McSherry 2013). Risk assessment tools typically focus on 'the main risk variables or factors associated with the risk of reoffending', with some exclusively assessing historical or '"static" risk factors that cannot change over time and that cannot be changed through treatment and intervention' (McSherry 2013: 45). Static factors may include the age of first offending and the offender's prior criminal history. Consequently, many assessments do not consider 'current clinical variables (such as response to treatment or motivation), protective factors (such as stable employment) or variables that reduce the risk of reoffending (such as physical illness or frailty)' (McSherry 2013: 45). AI-enabled approaches could – in theory – better accommodate some of the dynamic factors that exist in a person's life, for example by generating and analysing data on a multitude of contemporaneous factors in a person's life. As we will discuss further later, such approaches would seemingly require the disclosure of large volumes of private data about an individual, and the manner in which dynamic variables are weighted may be obscure and difficult to review or contest.

A second major criticism of forensic mental health risk assessment that McSherry identifies is the application of group data to individuals and the resulting likelihood of bias – a problem that would appear to be amplified by automated approaches based on ever larger, population-level datasets. Risk categorisation involves classifying an individual in relation to a group, such as 'high risk', 'medium risk' or 'low risk'. Yet assessors 'cannot say where, in this group, a given person lies and, therefore, cannot identify the precise risk an individual poses' (McSherry 2020: 30). In criminal law trials, this nuance may be missed by judges unless they receive expert witness guidance on how to interpret the results. Bias is also likely where risk assessment tools are developed based on one population and then used to assess risk in individuals of a different population. An example is a risk assessment tool developed using data from non-Indigenous offenders that is applied to Indigenous offenders. According to Kelly Hannah-Moffat (2013: 278), judges and other legal decision-makers may also mistake correlation and causation when actuarial risk tools are used in sentencing:

> 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.

Hannah-Moffat (2013) highlights a risk that courts and practitioners will struggle with the meaning of probabilistic scoring and overlook the substandard evidence of risk assessment tools.

Bias is a concern consistently raised in the digital context. Lorna McGregor (2023) has argued that automating risk assessment is unlikely to resolve the highly contested nature of risk prediction, particularly 'in the absence of

agreement on the factors that can be used to predict risk, and the exclusion of data points that may serve as a proxy for bias or discrimination'. It is now well-accepted that datasets that fuel automated systems, including those that utilise AI, can themselves incorporate or perpetuate bias in one or more respects. Datasets are created by humans, and choices made by humans about what data to collect in the first place, what data to include in training datasets, and how that data is organised or tagged can result in individual and systemic biases (conscious or unconscious), which become built into such systems. The link between actuarial prediction and automated tools is important here: automation in many cases relies on the same kinds of statistical analysis, so embeds concerns about the bias that McSherry and others allude to in earlier actuarial methods. As McSherry (2020) has indicated, dataset choice affects mathematical analysis whether by actuaries or computers.

A well-known example of such bias emerging in the criminal law context is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an automated decision-making tool used by judges in some US jurisdictions to inform sentencing decisions based on an assessment of an individual's risk of recidivism (Angwin et al. 2016). COMPAS was shown to produce a higher rate of 'false positive' results for black defendants than for white defendants; in other words, it was more likely to incorrectly predict that black people assessed by the system would re-offend when compared to results for white people (Angwin et al. 2016; cf. Flores and Bechtel 2016). The tool did not explicitly use race as a variable for predicting the risk of reoffending, but race correlated with variables that are themselves correlated with risk classification, such as poverty, joblessness and social marginalisation, producing biased results (Angwin et al. 2016). It is inevitable that similar forms of bias will emerge in experiments with algorithmic risk assessment and prediction in forensic mental health settings.

A third critique of actuarial risk assessment that McSherry (among others) has raised is specific to the use of AI and other automated systems; namely, that such systems tend to obfuscate, or at least lack transparency, in one or more senses (see Pasquale 2016). Such opacity may exist on several levels, including in the use of pretentious language, with technology vendors exaggerating or cultivating a mystique around technological capabilities (see Paterson 2022). Obfuscation also appears in the internal operation of AI-enabled systems, with some systems remaining inscrutable or hard to explain as to how conclusions are reached. There are several potential sources of this opacity, including the sheer scale of data being used and the high number of variables being used in AI and machine learning systems which entails 'a degree of unavoidable complexity' (Burrell 2016: 5); and intentional secrecy in service to proprietary interests, with private companies imposing confidentiality clauses and intellectual property protections over data and algorithms (see Slobogin 2021: 109–10). Questioning the outcomes produced by these tools – even on fundamental matters such as which of an individual's characteristics were taken into account in the tool's calculation of that person's level of risk, and how

variables were weighted – is made more difficult by the lack of transparency in the development and use of some partially or fully automated risk assessment tools. Another source of obfuscation may be a lack of explainability in the sense of the translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation by auditors, regulators, researchers and users and, related to this, limited technical literacy on the part of users/subjects/affected persons (Burrell 2016).

On this point about transparency, McSherry has suggested that '[t]he results generated and the methodology underlying risk assessment tools must be interpretable in order for appropriate judicial decisions to be made' (McSherry 2020: 36–37). Slobogin (2021: 111) more forcefully argues that '[e]ven if it turns out that advanced machine-learning [risk assessment instruments] are demonstrably more accurate than simpler versions (which is unlikely), they should be banned from criminal proceedings, at least when they are "inscrutable" '. Slobogin cites a fundamental compromise to fairness when litigants, policymakers and decision-makers are not provided with accurate and clear information about how the software works. Others have concluded that the use of predictive algorithms cannot be legitimised in sentencing decisions but have 'left open the door for legitimate use in treatment decisions' (van Eijk 2020: 119), a point to which we will return later in this chapter.

Concerns about transparency and avoiding 'black box' decision-making also tie into a persistent criticism of the way statistical methods and probabilistic reasoning have been used to give risk assessment the veneer of objectivity and certainty. A framing of algorithmic risk assessment as a 'technical' approach – whether using AI or not – conveys a view of a neutral, administrative and objective process for improving public safety (Rose 1998). A secondary effect of this projected certainty is that the stronger the public confidence in the accuracy of such assessments, misplaced though this may be, the more likely that 'moral outrage' will be attributed to public authorities who bear an implied culpability when certain types of tragedy occur, such as a person completing suicide or harming others during a mental health crisis (Szmukler and Rose 2013) – and the greater the social acceptability of and even public demand for intrusions on individuals deemed 'risky'. Digital forms of monitoring and assessment may amplify this misleading projection of facticity, and further pave the way for expanded preventive detention powers. As the SAFE-HOME example suggests, these intrusions are likely to disproportionately affect those with mental health conditions and psychosocial disabilities.

Such an 'allure of certainty' raises the possibility that automated risk assessment and management will be unreflectively normalised in preventive justice or psychiatric schemes, repeating harms of earlier opaque and misguided risk assessment approaches. Along these lines, Nikolas Rose (1998: 190–91) has offered the following critique of 'risk thinking' more broadly and the associated technologies and discourses of risk:

> Given that clinicians are not experts in the methodology and statistics 'black boxed' within risk-assessment scales, the populations upon

> which they are standardized, the generalization from one population to another, the moral and social judgments involved in decisions as to what does or does not count as dangerous behavior in the populations on which the scales were developed, the 'false positive rates', the effects of changes since the time the scales were constructed, the implications of national and cultural variation, what are the implications of their uncritical clinical acceptance of the 'objectivity' of such scales? . . . [T]he facticity conferred by scales and numbers, in which the decisions, calculations, techniques, and assumptions of the methods disappear into the apparent objectivity of the single number, not only serves to increase the appearance of accuracy. It also serves to decrease contestability and to imply specious pre-vision.

This critique was made in 1998 in terms that could well apply to concerns about the use of automated and AI-enabled risk assessment technologies in mental health settings a quarter of a century later. Yet, today, adding to this already perplexing array of factors for which to account, are ever-new forms of software, proprietary barriers to algorithmic methods and datasets, cheaper and more sophisticated monitoring and surveillance technologies and an increasingly interconnected communication ecosystem.

A final relevant point is that some mental health practitioners have raised ethical concerns about the pressure they face from courts to provide risk assessments for purposes such as continued supervision or detention after a sentence is complete (see e.g. Sullivan, Mullen and Pathé 2005). As McSherry outlines, this risk assessment role can leave clinicians feeling pressed to serve as 'agents of supervision, social control and monitoring' rather than as 'independent clinicians' (McSherry 2014: 787; see also Sullivan, Mullen and Pathé 2005; Gunn 2000). In the digital context, a related concern is the potential for automation to generate cheap, if limited, software and monitoring devices that will be attractive to system administrators who are drawn to a substitute for more expensive, expert and empathetic professionals (Pasquale 2020).

Indeed, with regards to the role of mental health experts, McSherry (2020) has observed that perhaps the most prominent strategy to constrain an unreflective reliance on algorithmic risk assessment, though one that remains contentious, is 'structured professional judgment'. Structured professional judgment combines statistical or actuarial risk prediction with clinical methods; in other words, clinicians use risk prediction tools alongside their professional judgment to make decisions or recommendations about a person's 'risk'. It was proposed to improve upon the *unstructured* decision-making by judges, parole authorities and mental health professionals described earlier, while mitigating the mechanistic, inflexible and opaque qualities of 'purely' algorithmic assessment. This proposition has been broadly convincing to both courts and in the forensic mental health field, and structured professional judgment 'has become an accepted forensic method to help identify those who are at low, moderate or high risk of harming others' (McSherry 2020: 23).

For McSherry (2020: 38), structured professional judgment seems to offer 'the "least worst" option' currently available, even as it remains contentious and raises 'serious ethical and human rights issues because of the consequences for the individuals concerned'. McSherry's work highlights the potential role of a human rights lens in identifying better than 'least worst' options for reform and change, to which we now turn.

## Human Rights: False Hope or a Targeted Mechanism of Accountability?

McSherry's work applies a human rights lens to the issues raised by digital technological developments in the mental health context. Although there is a growing body of work concerning automation and human rights more broadly (Access Now 2018; Amnesty International and Access Now 2018; McGregor, Murray and Ng 2019), little if any has provided an explicit focus on automation in the mental health context generally (for a notable exception, see Cosgrove et al. 2020), or the forensic mental health context in particular.

Researchers concerned with the consequences for humans and human society of AI and automated decision-making have tended *in general* to focus on 'ethical' approaches to such technologies, drawing on, for instance, bioethics or virtue ethics approaches (Fjeld et al. 2020), and ethical framing has been extended to the mental health context (e.g. Luxton 2015; Capon et al. 2016; Lederman et al. 2020; Martinez-Martin 2020; Martinez-Martin et al. 2020). Ethical principles for assessing, designing and utilising AI have also been proposed by industry, governments, international and intergovernmental organisations and professional associations as an appropriate means to guide the design, development and deployment of AI (see e.g. Berendt 2019; Jobin, Ienca and Vayena 2019; Fjeld et al. 2020). The proliferation of these non-binding codes, guidelines and recommendations for ethical AI has been attributed at least partly to industry resistance to stronger regulation of the design, development and use of AI and automated decision-making (United Nations General Assembly 2019). Such ethics statements typically identify a set of principles that should guide the development and use of AI to benefit humans, or at a minimum to avoid harm. While there is considerable variation in their content, they commonly include principles of transparency, fairness, non-maleficence, responsibility, privacy and trust, as well as concepts associated with social justice and human rights, such as equity, justice and access. The criticisms of forensic mental health risk assessment discussed in the preceding section map onto several of these principles, including issues of bias, transparency, privacy and trust, as well as justice related to the use of these tools in determining a person's 'riskiness'.

Ethical principles for AI provide some framework for assessing and responding to the issues raised by automated and AI-enabled risk assessments in the context of mental health. Yet assessing AI and automated decision-making on the basis of ethical principles has been criticised for several reasons. The most

relevant for present purposes is that statements of ethical AI principles tend to be articulated in overly broad terms without reference to normative underpinnings (Algorithm Watch 2019). In consequence, they are open to multiple interpretations (Pizzi, Romanoff and Engelhardt 2020) – including interpretations that may require a much lower standard than that established in human rights instruments (Smuha 2021) – and also to manipulation (Rességuier and Rodrigues 2020). This means 'there are serious problems of conceptual incoherence, conflicts among norms are rarely acknowledged, meaningful input is rarely sought from stakeholders and accountability mechanisms are absent' (United Nations General Assembly 2019: para 40; see also Yeung, Howes and Pogrebna 2020: 76–106).

A second criticism is that, while ethics-based statements often mention elements of accountability such as 'transparency' and 'responsibility', they do not tend to identify or prescribe mechanisms for enforcement of standards or consequences for breach (Fukuda-Parr and Gibson 2021: 42). This relates to broader criticisms of the origin of many statements of ethical AI principles in self-regulation efforts by the tech industry, which include allegations of 'ethics-washing' whereby Big Tech has avoided or delayed the imposition of binding forms of regulation while giving the appearance of being socially responsible (Wagner 2018; Hagendorff 2020).

Human rights principles, such as equality and non-discrimination, are sometimes mentioned in statements of ethical principles, or human rights is mentioned in general as one of a series of ethical principles for assessing the harms and benefits of AI (Fjeld et al. 2020; Fukuda-Parr and Gibbons 2021: 37). However, there has only recently been attention on the potential of a human rights approach to become the over-arching framework (rather than a subset of ethics) to address the limitations of ethics-based analyses and provide 'a sound normative framework to steer AI-systems towards the good' (Smuha 2021).

The main purported advantages of the human rights perspective are that it provides a set of internationally agreed-upon values and that those values are, at least to an extent, enforceable via laws, principles and standards whose interpretation and application is supported by existing enforcement mechanisms and a wealth of jurisprudence, research and other guidance (Latonero 2018; Berthet 2019; Aizenberg and van den Hoven 2020; Yeung, Howes and Pogrebna 2020; Smuha 2021). Fukuda-Parr and Gibbons (2021: 34–35) highlight several features of a human rights approach that set it apart from other statements of AI ethics, including its focus on structural issues and power imbalances; its imposition of obligations on States and other parties to ensure the realisation of rights (not to simply refrain from breaching them); and its underpinning by guiding principles including universality, equality and non-discrimination, participation, and accountability and remedy. These authors have also noted the growing attention on the human rights implications of technology from 'the UN human rights machinery', with special rapporteurs and other mandates 'call[ing] for studies and debates to develop norms, principles and standards' (Fukuda-Parr and Gibbons 2021: 34).

Criticisms have been levelled at human rights approaches to the assessment, design and deployment of AI and automated decision-making. Some of these engage long-standing critiques of human rights law, including that it is not enforceable in the conventional sense – the extent to which it is implemented and enforced varies greatly and depends in many states on international commitments (demonstrated via ratification of international instruments) being implemented in domestic law (see Yeung, Howes and Pogrebna 2020).

Two other criticisms are specific to human rights approaches to automated and AI-enabled technologies. The first is that existing recommendations for human rights-compliant AI tend not to appreciate or account for the technical complexities involved (Floridi 2010); the second is that attempts to articulate human rights requirements provide limited practical guidance for their implementation because they are stated at similar levels of generality to statements of AI ethics (Smuha 2021). Other critiques take aim at the foundations of human rights, including concerns that core tenets of humanism and liberalism in the era of Big Data not only fail to prevent the expansion of inequality and widespread social harm but potentially entrench it and work to guarantee the supremacy of private firms over individuals (see e.g. Benthall and Goldenfein 2020; McQuillan 2021).

Deeply engaging with these broad critiques is beyond the scope of this chapter. For our purposes, it is notable that there have been some recent efforts to develop practical human rights guidance in the context of AI and automated decision-making. These efforts have at least begun addressing concerns about the ambiguous practical applicability of human rights. For example, the UN Special Rapporteur for the Rights of Persons with Disabilities, Gerard Quinn, published a 2022 thematic study on artificial intelligence and its impact on persons with disabilities (Human Rights Council 2021). Quinn observed that 'there has been little detailed assessment of the direct benefits and potential harms of artificial intelligence for the world's approximately 1 billion persons with disabilities' (Human Rights Council 2021: 6). He concluded that 'a fundamental reset of the debate [about artificial intelligence] is needed, based on more evidence and greater consideration of the rights and obligations contained in the Convention on the Rights of Persons with Disabilities and other human rights instruments' (Human Rights Council 2021: 17).

McSherry's leading steps to apply human rights to contemporary developments in forensic mental health risk assessment point to several of the advantages, and limitations, of a human rights approach to the issues discussed earlier. Of primary concern for McSherry has been the human right to liberty, which she asserts is clearly violated by risk assessment tools that are used to justify and facilitate indefinite and preventive detention (McSherry 2013, 2020). McSherry (2017: 69) has argued that 'human rights may be able to provide a framework for change despite the intractable nature of risk as a criterion for detention of those with mental impairment' and has discussed this approach specifically in relation to predictive algorithms for preventive detention (see also McSherry 2020).

Key to any such change is enforceability of human rights protections, which is variable at domestic levels throughout the world. McSherry (2020: 30) notes the tendency for community protection arguments that support the use of preventive detention to trump rights arguments in countries such as Australia and New Zealand, which have 'relatively weak rights protection', even where preventive detention regimes have been declared to constitute human rights violations by the Human Rights Council. In contrast, limits have been placed on preventive detention schemes in states with stronger human rights protections, including challenges to such a scheme in Germany which resulted in the Federal Constitutional Court declaring preventive detention orders unconstitutional on the basis that they violated the right to liberty (McSherry 2020: 34–36). McSherry acknowledges that these protections are not absolute, despite preventive detention constituting a 'clear' violation of the right to liberty, which touches on criticisms of human rights law lacking enforceability in many places. We would nevertheless agree with McSherry's (2020: 36) assertion that '[j]urisprudence having an influence rather than being binding is still better than having no influence at all'.

To date, McSherry's work has not specifically engaged with the way that automated and AI-enabled risk assessment tools might be utilised to *promote* and *protect* the rights of accused and convicted persons. Raso and colleagues (2018: 20), for example, suggest that certain risk assessment tools, including those that use machine learning, could actually strengthen people's right to liberty where their use results in 'low-risk' individuals benefitting from greater pre-trial release and shorter sentences, with the wider community also benefitting from a lower crime rate. A US policy simulation study of pre-trial detention practices using New York data, for example, concluded that 42% more arrestees would be released as 'low risk' if a risk assessment instrument using machine learning replaced cash bail (Kleinberg et al. 2017). Slobogin (2021: fn 166) points to numerous other studies indicating that jurisdictions using risk assessment instruments can release a greater number of people without raising crime rates. He concludes that overall, when properly regulated, with peer review and vetting, training for judges, lawyers and correctional officials, and a willingness to abandon the measures if hypothesised benefits fall short, risk assessment tools can be a crucial means of safely and humanely dismantling the massive jail and prison complex in the US (Slobogin 2021: 158–63). This is arguably an optimistic picture, and one that appears to take specific aim at the anomalously high rates of US incarceration. Further, Slobogin adds the caveat that if after all those steps of rigorous peer review and auditing were taken, and the scheme was found not to have achieved its aims, that it probably ought to be abandoned (Slobogin 2021: 163). Regardless, such a pragmatic approach does not negate the rights concerns that these technologies raise to those affected, particularly those deemed 'high risk' and others whose subjection to risk assessment results in longer detention or more intensive monitoring.

## Potential Avenues for Research on Rights and Algorithmic Risk Assessment

If one adopts the human rights framework, there is clear scope to extend McSherry's focus on the right to liberty to other human rights issues raised by automated and AI-enabled risk assessments, including the rights articulated in the International Covenant on Civil and Political Rights (ICCPR) and Convention on the Rights of Persons with Disabilities (CRPD). This could include rights concerning equality before the law and non-discrimination, health, protection from arbitrary arrest and detention, rights to a fair hearing and the presumption of innocence, and (perhaps particularly in the light of increased technological capacity for remote monitoring and surveillance) rights to protection from interference with privacy and family life.

Gerard Quinn, in his aforementioned report, argued that equality and non-discrimination must be the primary considerations when assessing the rights implications of AI for disabled people (Human Rights Council 2021). The possibility of bias in the training data for risk assessment tools clearly raises concerns about the violation of this right, although arguments that some digital tools may reduce racial and other disparities in bail and sentencing raise the possibility that some technologies may actually enhance the protection of this right (Raso et al. 2018: 20).

The possibility of individuals being erroneously classified as 'high-risk', and subsequently detained on that basis, or the use of risk assessment technology where the basis on which individuals have been assessed as high risk and detained is not transparent or open to appeal, raise questions about such individuals' enjoyment of the right to be free from arbitrary arrest and detention (ICCPR: Article 9). The combination of proprietary interests, inherent complexity and inscrutability of automated systems makes it especially difficult to challenge algorithmic determinations in court settings and also raises concerns in relation to the rights to a fair hearing and to be presumed innocent until found guilty (ICCPR: Article 14; CRPD: Article 14). Automated risk assessment systems rely on the generation, storage and analysis of large and even vast amounts of personal data, raising significant privacy concerns, which are acute in the context of 'passive monitoring' and surveillance. Such matters may constitute violations of the freedom from arbitrary or unlawful interference with privacy, family, home and correspondence articulated in the ICCPR (Article 17) and the CRPD (Article 22). In this sense, it is possible that measures to promote 'fairness' or human rights compliance in one respect will have contrary consequences in another – such as measures to improve the quality of personal data used to train automated systems, which may result in less biased systems but may have deleterious consequences for individuals' privacy.

The right to the highest attainable quality of physical and mental health on an equal basis with others may also be undermined by data-driven evaluation of individuals, including through monitoring and surveillance (CRPD:

Article 25). Dainius Pūras, then UN Special Rapporteur on the Right to the Highest Attainable Physical and Mental Health, has warned (albeit briefly) of the negative impact on the right to health of expanding surveillance technologies that 'categorize an individual for commercial, political or additional surveillance purposes' (Human Rights Council 2020: para 75). An example might be involuntary 'monitoring conditions' on people detained in forensic psychiatric settings in the form of electronic ankle bracelets, which proceed against the submissions of medical practitioners, and in a manner that hinders the rehabilitation and treatment of the individuals concerned (see e.g. Miller 2015). Others have argued that the very designation of people as 'high risk' through predictive algorithmic approaches to sentencing could violate rights-based protection of human dignity given that it impacts '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' (Ward 2011: 106).

Further work is clearly required to move from this kind of general discussion of human rights implications to detailed consideration in the mental health context and in relation to specific technologies such as forensic risk assessments. Fruitful exploration may include close analysis of the applicability of existing mechanisms for the realisation and enforcement of each of the above human rights to specific applications of AI and automation, and consideration of the implementation of AI-specific human rights measures proposed by bodies such as the European Commission's AI High-Level Expert Group, the UN Special Rapporteur on the Rights of Persons with Disabilities and the Australian Human Rights Commission, such as human rights impact assessments, auditability requirements, ex-ante oversight, stakeholder consultation and procedures for redress (High-Level Expert Group on Artificial Intelligence 2019; Human Rights Council 2021; Australian Human Rights Commission 2021).

Cross-disciplinary and inter-disciplinary work, which brings together mental health, legal and computer science scholars, including those working from a lived experience of both mental health conditions and detention or supervision in forensic mental health systems, may be necessary to investigate the utility and implementation of these and other measures in relation to forensic mental health risk assessment.

There has been some commentary from people with lived experience of mental health conditions or psychosocial disabilities, among which prominent concerns include: promoting a 'right to explanation' concerning algorithmic decision-making for individuals (both the right of an individual to understand how a decision about them using algorithmic technology was made, but also to query the values that go into a particular algorithmic decision system) (Carr 2020); the risk of discrimination or harm where sensitive personal information is leaked, stolen, sold or scraped from social media and other 'public' fora (Consumers of Mental Health Western Australia 2018); and concerns about the deployment of data-driven technologies in coercive psychiatric

interventions and in policing (Harris 2019; Carr 2020). However, such commentaries are scarce, and there seems to be a general marginalisation of lived experience perspectives from research concerning algorithmic and data-driven approaches to mental health (Gooding and Kariotis 2021).

As a final point, the aim or even mandate to elevate the subjective standpoint of disabled persons, which is incorporated throughout the CRPD, also appears to run counter to some of the theoretical aims behind automated prediction tools, offering important avenues for future work. Jackie Leach Scully and Georgia Van Toorn (2021) have argued that broader 'datafication' of the human body will delineate increasingly rigid boundaries between normality and disability. This impulse to quantify and distinguish embodied difference, they argue, 'diverts attention from the realities of disabled lives, at a time when disability scholars and activists are arguing for more rather than less attention to the lived experience of disability' (Scully and Van Toorn 2021).

One expression of this attention to the lived experience of disability in the forensic mental health context, is the call for 'collaborative risk assessment and management' in forensic mental health settings. Sarah Markham (2020: 5) suggests that collaborative risk assessment and management requires 'a more fluid and responsive culture with increased emphasis on relational safety and epistemic regard for patient self-insight and testimony'. Although collaborative risk assessment is outside the scope of this chapter, it seems another potentially generative line of enquiry in broader discussions about risk assessment today, given such an approach appears to be antithetical to the aim of using passive monitoring and Big Data analytics to 'overcome the limitations' of the subjective accounts of patients/service users and clinicians (see also Markham 2021).

Another possibility, given the ambivalent impact of the automated and AI-enabled approaches to risk assessment discussed in this chapter, is that it may be more productive to get past the binary idea that either computers can improve on the fallibility of human decisions, or that humans can improve on the fallibility of computer decisions. Instead, it may be more helpful to acknowledge that what lies beneath these debates are longstanding and intractable tensions in law of rules and equality on the one hand, and discretion and individuality on the other (Goldenfein 2019). Structured professional judgment may well have to remain the 'least worst option' for mediating these tensions (McSherry 2020: 38), which next invites questions of (1) what makes a good professional judgment or decision; and (2) what role automated and AI-enabled approaches might play in making them. Human rights clearly have something to offer in this respect.

As a final point, the broader critique of a human rights approach to automation and AI warrants continued attention. Taken seriously, such critiques could help to either refine rights-based efforts, where critical viewpoints are addressed within a human rights framework, or divert those who find the criticisms convincing to seek strategies beyond what human rights would seem to offer.

## Conclusion

The digital, data-informed evaluation of persons is advancing. The world is at the beginning of a long project of critiquing and regulating AI and other algorithmic technologies and understanding the role they play in attempts to make individuals more 'calculable'. Risk assessment in the forensic mental health context is one such area in which the legal and regulatory landscape is likely to change considerably in coming years – particularly in light of efforts to integrate biometric monitoring and surveillance capabilities of increasingly ubiquitous sensor technologies. The Trump administration's serious consideration of the 'SAFEHOME' proposal seems likely to be a portent of future efforts along these lines. Further study of these developments and their profound implications for people's human rights could contribute to efforts to guide the design, use, regulation and broad governance of AI and automation in this field more generally. Bernadette McSherry's leading scholarship provides a rich corpus from which to build this work.

## References

Access Now (2018) *Human Rights in the Age of Artificial Intelligence*. Available www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf (accessed 15 December 2022).

Aizenberg, E. and van den Hoven, J. (2020) 'Designing for Human Rights in AI', *Big Data & Society, 7*(2): 1–14.

Algorithm Watch (2019) *No Red Lines: Industry Defuses Ethics Guidelines for Artificial Intelligence*. Available https://algorithmwatch.org/en/industry-defuses-ethics-guidelines-for-artificial-intelligence/ (accessed 15 December 2022).

Amnesty International and Access Now (2018) *The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems*. Available www.torontodeclaration.org/declaration-text/english/ (accessed 15 December 2022).

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) 'Machine Bias', *ProPublica*, 23 May. Available www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=TiqCeZIj4uLbXl91e3wM2PnmnWbCVOvS (accessed 15 December 2022).

Australian Human Rights Commission (2021) *Human Rights and Technology – Final Report*. Available https://tech.humanrights.gov.au/sites/default/files/2021-05/AHRC_RightsTech_2021_Final_Report.pdf (accessed 15 December 20220).

Bartels, L. and Martinovic, M. (2017) 'Electronic Monitoring: The Experience in Australia', *European Journal of Probation*, 9(1): 80–102.

Benthall, S. and Goldenfein, J. (2020) 'Data Science and the Decline of Liberal Law and Ethics', *Social Science Research Network*. Available https://doi.org/10.2139/ssrn.3632577 (accessed 15 December 2022).

Berendt, B. (2019) 'AI for the Common Good?! Pitfalls, Challenges, and Ethics Pen-testing', *Paladyn: Journal of Behavioral Robotics*, 10(1): 44–65.

Berthet, A. (2019) *Why Do Emerging AI Guidelines Emphasize 'Ethics' over Human Rights?* OpenGlobalRights. Available www.openglobalrights.org/why-do-emerging-ai-guidelines-emphasize-ethics-over-human-rights/ (accessed 15 December 2022).

Burrell, J. (2016) 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', *Big Data & Society*, 3(1): 1–12.

Capon, H., Hall, W., Fry, C. and Carter, A. (2016) 'Realising the Technological Promise of Smartphones in Addiction Research and Treatment: An Ethical Review', *International Journal of Drug Policy*, 36: 47–57.

Carr, S. (2020) ' "AI Gone Mental": Engagement and Ethics in Data-driven Technology for Mental Health', *Journal of Mental Health*, 29(2): 125–30.

Castel, R. (1991) 'From Dangerousness to Risk', in G. Burchell, C. Gordon and P. Miller (eds) *The Foucault Effect: Studies in Governmentality*, Chicago: University of Chicago Press, 281–98.

Consumers of Mental Health Western Australia (2018) *9 Days Left to Make an Informed Decision about My Health Record*, 6 November. Available https://comhwa.org.au/blog/9-days-left-to-make-an-informed-decision-about-my-health-record (accessed 15 December 2022).

Cosgrove, L., Karter, J., Mcginley, M. and Morrill, Z. (2020) 'Digital Phenotyping and Digital Psychotropic Drugs: Mental Health Surveillance Tools That Threaten Human Rights', *Health and Human Rights Journal*, 22(2): 33–40.

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York: St Martin's Press.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020) 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI', *Social Science Research Network*. Available https://doi.org/10.2139/ssrn.3518482 (accessed 15 December 2022).

Flores, A.W. and Bechtel, K. (2016) 'False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks" ', *Federal Probation*, 80(2): 38–46.

Floridi, L. (2010) 'Information Ethics', in L. Floridi (ed) *The Cambridge Handbook of Information and Computer Ethics*, Cambridge: Cambridge University Press; Cambridge Core, 77–98.

Fukuda-Parr, S. and Gibbons, E. (2021) 'Emerging Consensus on "Ethical AI": Human Rights Critique of Stakeholder Guidelines', *Global Policy*, 12(S6): 32–44.

Goldenfein, J. (2019) 'Algorithmic Transparency and Decision-Making Accountability: Thoughts for Buying Machine Learning Algorithms', in S. Bluemmel (ed) *Closer to the Machine: Technical, Social, and Legal Aspects of AI*, Melbourne: Office of the Victorian Information Commissioner, 41–60.

Gooding, P. and Kariotis, T. (2021) 'Ethics and Law in Research on Algorithmic and Data-Driven Technology in Mental Health Care: Scoping Review', *JMIR Mental Health*, 8(6): e24668.

Gooding, P. and Resnick, K. (2020) 'Psychiatry and Law in the Digital Age: Untangling the Hype, Risk and Promise', *International Journal of Law and Psychiatry*, 70: 101553.

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Öhler, S., Tröster, G., Mayora, O., Haring, C. and Lukowicz, P. (2015) 'Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients', *IEEE Journal of Biomedical and Health Informatics*, 19(1): 140–48.

Gunn, J. (2000) 'Future Directions for Treatment in Forensic Psychiatry', *The British Journal of Psychiatry*, 176(4): 332–38.

Hagendorff, T. (2020) 'The Ethics of AI Ethics: An Evaluation of Guidelines', *Minds and Machines*, 30(1): 99–120.

Hannah-Moffat, K. (2013) 'Actuarial Sentencing: An "Unsettled" Proposition', *Justice Quarterly*, 30(2): 270–96.

Harris, L. (2019) 'The Rise of the Digital Asylum', *Mad In America*, 15 September. Available www.madinamerica.com/2019/09/the-rise-of-the-digital-asylum/ (accessed 15 December 2022).

High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines for Trustworthy AI*, European Commission. Available https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 15 December 2022).

Holmes, A. (2013) 'Is Risk Assessment the New Clinical Model in Public Mental Health?' *Australasian Psychiatry: Bulletin of Royal Australian and New Zealand College of Psychiatrists*, 21(6): 541–44.

Human Rights Council (2020) *Report of the Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health*, UN Doc A/HRC/44/48. Available www.ohchr.org/en/documents/thematic-reports/ahrc4448-right-everyone-enjoyment-highest-attainable-standard-physical (accessed 15 December 2022).

Human Rights Council (2021) *Report of the Special Rapporteur on the Rights of Persons with Disabilities*, UN Doc A/HRC/49/52. Available www.ohchr.org/en/documents/thematic-reports/ahrc4952-artificial-intelligence-and-rights-persons-disabilities-report (accessed 15 December 2022).

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The Global Landscape of AI Ethics Guidelines', *Nature Machine Intelligence*, 1(9): 389–99.

Kak, A. (2021) *Regulating Biometrics: Global Approaches and Urgent Questions*, AI Now Institute. Available https://ainowinstitute.org/regulatingbiometrics.pdf (accessed 15 December 2022).

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017) *Human Decisions and Machine Predictions*, NBER Working Paper Series No. 23180. Available www.cs.cornell.edu/home/kleinber/w23180.pdf (accessed 15 December 2022).

Latonero, M. (2018) *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data & Society. Available https://datasociety.net/library/governing-artificial-intelligence/ (accessed 15 December 2022).

Le, D.V., Montgomery, J., Kirkby, K.C. and Scanlan, J. (2018) 'Risk Prediction Using Natural Language Processing of Electronic Mental Health Records in an Inpatient Forensic Psychiatry Setting', *Journal of Biomedical Informatics*, 86: 49–58.

Lederman, R., D'Alfonso, S., Rice, S., Coghlan, S., Wadley, G. and Alvarez-Jimenez, M. (2020) *Ethical Issues in Online Mental Health Interventions*, ECIS 2020 Research Papers. Available https://aisel.aisnet.org/ecis2020_rp/66/ (accessed 15 December 2022).

Lupton, D. (2013) *Risk*, 2nd edn, Abingdon: Routledge.

Luxton, D.D. (2015) *Artificial Intelligence in Behavioral and Mental Health Care*, London: Academic Press.

McGregor, L. (2023) *Detention and Its Alternatives under International Law*, Oxford: Oxford University Press.

Markham, S. (2020) 'Collaborative Risk Assessment in Secure and Forensic Mental Health Settings in the UK', *General Psychiatry*, 33(5): e100291.

Markham, S. (2021) 'The Omnipresence of Risk and Associated Harms in Secure and Forensic Mental Health Services in England and Wales', *Social Theory & Health*. Available https://link.springer.com/article/10.1057/s41285-021-00167-z (accessed 15 December 2022).

Marks, M. (2019) *Artificial Intelligence Based Suicide Prediction*, Social Science Research Network. Available https://papers.ssrn.com/abstract=3324874 (accessed 15 December 2022).

Marks, M. (2020) 'Algorithmic Disability Discrimination', in A. Silvers, C. Shachar, I.G. Cohen and M.A. Stein (eds) *Disability, Health, Law, and Bioethics*, Cambridge: Cambridge University Press, 242–54.

Martinez-Martin, N. (2020) 'Chapter Three – Trusting the Bot: Addressing the Ethical Challenges of Consumer Digital Mental Health Therapy', in I. Bárd and E. Hildt (eds) *Developments in Neuroethics and Bioethics*: *Ethical Dimensions of Commercial and DIY Neurotechnologies*, Cambridge, MA: Academic Press, 69–91.

Martinez-Martin, N., Dasgupta, I., Carter, A., Chandler, J. A., Kellmeyer, P., Kreitmair, K., Weiss, A. and Cabrera, L.Y. (2020) 'Ethics of Digital Mental Health During COVID-19: Crisis and Opportunities', *JMIR Mental Health*, 7(12): e23776.

Mathiesen, T. and Rutherford, A. (2006) *Prison on Trial*, 3rd edn, Winchester: Waterside Press.

McGregor, L., Murray, D. and Ng, V. (2019) 'International Human Rights Law as a Framework for Algorithmic Accountability', *International & Comparative Law Quarterly*, 68(2): 309–43.

McQuillan, D. (2018) 'Mental Health and Artificial Intelligence: Losing Your Voice', *openDemocracy*. Available www.opendemocracy.net/en/digitaliberties/mental-health-and-artificial-intelligence-losing-your-voice-poem/ (accessed 15 December 2022).

McQuillan, D. (2021) 'Post-Humanism, Mutual Aid', in P. Verdegem (ed) *AI for Everyone? Critical Perspectives*, London: University of Westminster Press, 67–83.

McQuillan, D. (2022) *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*, Bristol: Bristol University Press.

McSherry, B. (2013) *Managing Fear: The Law and Ethics of Preventive Detention and Risk Assessment*, New York: Routledge.

McSherry, B. (2014) 'Throwing Away the Key: The Ethics of Risk Assessment for Preventive Detention Schemes', *Psychiatry, Psychology and Law*, 21(5): 779–90.

McSherry, B. (2017) 'Preventive Justice, Risk of Harm and Mental Health Laws', in T. Tulich, S. Bronitt and S. Murray (eds) *Regulating Preventive Justice*, New York: Routledge, 61–74.

McSherry, B. (2018) 'Computational Modelling, Social Media and Health-Related Datasets: Consent and Privacy Issues', *Journal of Law and Medicine*, 25(4): 894–98.

McSherry, B. (2020) 'Risk Assessment, Predictive Algorithms and Preventive Justice', in J. Pratt and J. Anderson (eds) *Criminal Justice, Risk and the Revolt against Uncertainty*, Cham: Springer International Publishing, 17–42.

McSherry, B. and Freckelton, I. (eds) (2013) *Coercive Care: Rights, Law and Policy*, Abingdon: Routledge.

McSherry, B. and Keyzer, P. (eds) (2011) *Dangerous People: Policy, Prediction, and Practice*, New York: Routledge.

Miller, P. and Rose, N. (1995) 'Production, Identity, and Democracy', *Theory and Society*, 24(3): 427–67.

Miller, S. (2015) 'The Use of Monitoring Conditions (GPS Tracking Devices)Re CMX [2014] QMHC 4', *Psychiatry, Psychology and Law*, 22(3): 321–26.

Mossman, D. (2009) 'The Imperfection of Protection Through Detection and Intervention', *Journal of Legal Medicine*, 30(1): 109–40.

Office of the Privacy Commissioner of Canada (2017) *Disclosure of Information about Complainant's Attempted Suicide to US Customs and Border Protection Not*

*Authorized Under the Privacy Act*. Available www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-federal-institutions/2016–17/pa_20170419_rcmp/ (accessed 15 December 2022).

Organisation for Economic Co-operation and Development (OECD) (2019) *Artificial Intelligence in Society*, Paris: OECD Publishing.

Pasquale, F. (2016) *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, MA: Harvard University Press.

Pasquale, F. (2020) *New Laws of Robotics: Defending Human Expertise in the Age of AI*, Cambridge, MA: Belknap Press of Harvard University Press.

Paterson, J.M. (2022) 'Misleading AI: Regulatory Strategies for Algorithmic Transparency in Technologies Augmenting Consumer Decision-Making', *Loyola Consumer Law Review*, 34(SI): 558–81.

Paterson, J.M. and Maker, Y. (2021) *AI in the Home: Artificial Intelligence and Consumer Protection*, Social Science Research Network. Available https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3973179 (accessed 15 December 2022).

Pizzi, M., Romanoff, M. and Engelhardt, T. (2020) 'AI for Humanitarian Action: Human Rights and Ethics', *International Review of the Red Cross*, 102(913): 145–80.

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C. and Levin, K. (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks*, Berkman Klein Center for Internet & Society Research Publication. Available https://dash.harvard.edu/handle/1/38021439 (accessed 15 December 2022).

Resnick, K.S. and Appelbaum, P.S. (2019) 'Passive Monitoring of Mental Health Status in the Criminal Forensic Population', *The Journal of the American Academy of Psychiatry and the Law*, 47(4): 457–66.

Rességuier, A. and Rodrigues, R. (2020) 'AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics', *Big Data & Society*, 7(2): 1–5.

Rose, N. (1998) 'Governing Risky Individuals: The Role of Psychiatry in New Regimes of Control', *Psychiatry, Psychology and Law*, 5(2): 177–95.

Scully, J.L. and Van Toorn, G. (2021) *Datafying Disability: Ethical Issues in Automated Decision Making and Related Technologies – AABHL 2021*, 19 November. Available www.aabhlconference.com/3563 (accessed 15 December 2022).

Slemon, A., Jenkins, E. and Bungay, V. (2017) 'Safety in Psychiatric Inpatient Care: The Impact of Risk Management Culture on Mental Health Nursing Practice', *Nursing Inquiry*, 24(4): e12199.

Slobogin, C. (2012) 'Risk Assessment', in J. Petersilia and K.R. Reitz (eds) *The Oxford Handbook of Sentencing and Corrections*, New York: Oxford University Press, 196–214.

Slobogin, C. (2021) *Just Algorithms: Using Science to Reduce Incarceration and Inform a Jurisprudence of Risk*, Cambridge: Cambridge University Press.

Smuha, N.A. (2021) 'Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea', *Philosophy & Technology*, 34(1): 91–104.

Stevenson, M.T. and Doleac, J.L. (2019) *Algorithmic Risk Assessment in the Hands of Humans*, Social Science Research Network. Available https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440 (accessed 15 December 2022).

Sullivan, D.H., Mullen, P.E. and Pathé, M.T. (2005) 'Legislation in Victoria on Sexual Offenders: Issues for Health Professionals', *The Medical Journal of Australia*, 183(6): 318–20.

Szmukler, G. and Rose, N. (2013) 'Risk Assessment in Mental Health Care: Values and Costs', *Behavioral Sciences & the Law*, 31(1): 125–40.

Travis, S. (2019) 'Florida Wants to Amass Reams of Data on Students' Lives', *Sun-Sentinel.Com*, 9 July. Available www.sun-sentinel.com/local/broward/parkland/florida-school-shooting/fl-ne-school-shooting-database-deadline-20190709-i4oc-smqeivdmrhpauhyaplg52u-story.html (accessed 15 December 2022).

Trinhammer, M.L., Merrild, A.C.H., Lotz, J.F. and Makransky, G. (2022) 'Predicting Crime During or After Psychiatric Care: Evaluating Machine Learning for Risk Assessment Using the Danish Patient Registries', *Journal of Psychiatric Research*, 152: 194–200.

United Nations General Assembly (2019) *Report of the Special Rapporteur on Extreme Poverty and Human Rights 11 October*, UN Doc A/74/493. Available www.ohchr.org/en/documents/thematic-reports/a74493-digital-welfare-states-and-human-rights-report-special-rapporteur (accessed 15 December 2022).

Valentine, L., D'Alfonso, S. and Lederman, R. (2022) 'Recommender Systems for Mental Health Apps: Advantages and Ethical Challenges', *AI & SOCIETY*. Available https://pubmed.ncbi.nlm.nih.gov/35068708/ (accessed 15 December 2022).

van Eijk, G. (2020) 'Algorithmic Reasoning: The Production of Subjectivity Through Data', in R. Peeters and M. Schuilenburg (eds) *The Algorithmic Society*, Abingdon: Routledge, 119–34.

Wagner, B. (2018) 'Ethics as An Escape from Regulation. From "Ethics-Washing" To Ethics-Shopping?' in E. Bayamlioğu, I. Baraliuc, L. Janssens and M. Hildebrandt (eds) *Ethics as An Escape from Regulation: From "Ethics-Washing" To Ethics-Shopping?* Amsterdam: Amsterdam University Press, 84–89.

Wan, W. (2019) 'White House Weighs Controversial Plan on Mental Illness and Mass Shootings', *Washington Post*, 9 September. Available www.washingtonpost.com/health/white-house-considers-controversial-plan-on-mental-illness-and-mass-shooting/2019/09/09/eb58b6f6-ce72–11e9–87fa-8501a456c003_story.html (accessed 15 December 2022).

Ward, T. (2011) 'Human Rights and Dignity in Offender Rehabilitation', *Journal of Forensic Psychology Practice*, 11(2–3): 103–23.

Whittaker, M., Alper, M., Bennett, C.L., Hendren, S., Kaziunas, L., Mills, M., Morris, M.R., Rankin, J., Rogers, E., Salas, M. and West, S.M. (2019) *Disability, Bias, and AI*, AI Now. Available https://ainowinstitute.org/disabilitybiasai-2019.pdf (accessed 15 December 2022).

Yeung, K., Howes, A. and Pogrebna, G. (2020) 'AI Governance by Human Rights Centred-Design, Deliberation and Oversight: An End to Ethics Washing', in M.D. Dubber, F. Pasquale and S. Das (eds) *The Oxford Handbook of Ethics of AI*, New York: Oxford University Press, 76–106.

Zalnieriute, M., Bennett Moses, L. and Williams, G. (2019) 'The Rule of Law and Automation of Government Decision-Making', *Modern Law Review*, 82(3): 425–55.