REVIEW
published: 17 March 2020
doi: 10.3389/fpsyg.2020.00220

Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective

Leda Tortora[1]*, *Gerben Meynen*[2,3], *Johannes Bijlsma*[2], *Enrico Tronci*[4] and *Stefano Ferracuti*[1]

[1] Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy, [2] Willem Pompe Institute for Criminal Law and Criminology/Utrecht Centre for Accountability and Liability Law (UCALL), Utrecht University, Utrecht, Netherlands, [3] Faculty of Humanities, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [4] Department of Computer Science, Sapienza University of Rome, Rome, Italy

Advances in the use of neuroimaging in combination with A.I., and specifically the use of machine learning techniques, have led to the development of brain-reading technologies which, in the nearby future, could have many applications, such as lie detection, neuromarketing or brain-computer interfaces. Some of these could, in principle, also be used in forensic psychiatry. The application of these methods in forensic psychiatry could, for instance, be helpful to increase the accuracy of risk assessment and to identify possible interventions. This technique could be referred to as 'A.I. neuroprediction,' and involves identifying potential neurocognitive markers for the prediction of recidivism. However, the future implications of this technique and the role of neuroscience and A.I. in violence risk assessment remain to be established. In this paper, we review and analyze the literature concerning the use of brain-reading A.I. for neuroprediction of violence and rearrest to identify possibilities and challenges in the future use of these techniques in the fields of forensic psychiatry and criminal justice, considering legal implications and ethical issues. The analysis suggests that additional research is required on A.I. neuroprediction techniques, and there is still a great need to understand how they can be implemented in risk assessment in the field of forensic psychiatry. Besides the alluring potential of A.I. neuroprediction, we argue that its use in criminal justice and forensic psychiatry should be subjected to thorough harms/benefits analyses not only when these technologies will be fully available, but also while they are being researched and developed.

Keywords: neuroprediction, artificial intelligence, recidivism, forensic psychiatry, risk assessment, neurolaw

OPEN ACCESS

Edited by:
José M. Muñoz,
Universidad Europea de Valencia, Spain

Reviewed by:
Vaughn R. Steele,
Yale University, United States
Cristina Scarpazza,
University of Padova, Italy

****Correspondence:***
Leda Tortora
leda.tortora@hotmail.it

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *24 August 2019*
Accepted: *31 January 2020*
Published: *17 March 2020*

Citation:
Tortora L, Meynen G, Bijlsma J, Tronci E and Ferracuti S (2020) Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective. Front. Psychol. 11:220. doi: 10.3389/fpsyg.2020.00220

INTRODUCTION

Risk assessment is a crucial component of the criminal justice system. In recent years, there has been a growing interest in the development of new tools and techniques to improve risk assessment in the field of forensic psychiatry and criminal justice (Monahan and Skeem, 2015). Currently, more than 200 violence risk assessment tools, often integrated clinical-actuarial instruments, have been developed to predict violent, antisocial, and sexual behavior (Singh et al., 2014), and their use seems to be vastly increasing in criminal justice settings (Conroy and Murrie, 2007).

The central aim of these methods is to identify high-risk and low-risk offenders correctly. Depending on the jurisdiction, they are used to inform a range of medico-legal decisions, for instance regarding sentencing, parole, civil commitment, death penalty, disposition in juvenile courts, and discharge following findings of insanity (Conroy and Murrie, 2007). In recent years, A.I. (Artificial Intelligence) is being used to enhance the predictive accuracy of risk assessment.

The use of algorithmic risk assessment has grown along with the research in the field of neuroimaging, leading to the development of 'brain-reading' techniques that are, to some limited extent, able to decode mental states based on a person's brain activity (Haynes and Rees, 2006), or to classify people in groups based on their brain structure and functionality (Koutsouleris et al., 2012). A possible forensic application of the technique is to identify dangerous offenders. The combination of A.I. and neuroimaging has led to the development of what can be called 'A.I. neuroprediction,' which is the use of structural or functional brain parameters coupled with machine learning methods to make clinical or behavioral predictions. Perhaps, in the near future, A.I. neuroprediction could be more generally used to predict the risk of recidivism in forensic psychiatry and criminal justice. However, application of such techniques raises legal and ethical issues.

The purpose of this paper is to identify possibilities and challenges regarding the possible future use of A.I. neuroprediction of violence and recidivism in the fields of forensic psychiatry and criminal justice, discussing legal implications and ethical issues. In the next section, we will discuss risk-assessment techniques. In the third section, we consider current 'brain-reading' techniques that use neuroimaging coupled with A.I. In the fourth section, we provide an overview of recent neuroprediction studies using neuroimaging data coupled with A.I. to predict recidivism. In the fifth section, we discuss technological limitations and pitfalls of predictive analysis. Finally, in the sixth section, we discuss the ethical and legal issues raised by the application of these techniques.

RISK ASSESSMENT: THE STATE OF THE ART

In the past two decades, in both the US and Europe, interest in and research on violence risk assessment tools have significantly increased, providing different approaches varying from strictly actuarial tools, based on regression, to algorithmic risk assessment, providing a probabilistic estimate of reoffending, to structured professional judgment (Hart, 1998; Douglas and Kropp, 2002). Initially, actuarial methods dominated the field, but their predictive value remained quite limited, if not disappointing (Fazel et al., 2012).

Risk variables associated with an increased likelihood of an individual acting violently or aggressively include criminogenic needs (individual characteristics that increase the risk of recidivism), demographics, socioeconomic status, and intelligence (Gendreau et al., 1996). Risk factors are typically divided into static factors, that are historical and do not change (e.g., criminal history, offense types, childhood abuse) and dynamic factors that are, in principle, changeable and therefore they provide the opportunity for intervention, modifying future risk (e.g., impulsivity, drug use, social support, job, compliance with treatment). Some dynamic factors are quite stable, while others are more "fluid." Dynamic factors need to be measured multiple times, sometimes within short intervals.

At present, the results of risk assessment tools, however, are far from perfect, especially for long term prediction; current criminal risk assessment tools show poor to moderate accuracy, and a good balance between false positives and false negatives is an issue that should be considered, depending both on the social and political context and on the stage of the criminal justice process in which the tool is used (Douglas et al., 2017). Generally, when a risk assessment tool classifies an individual as low-risk, it is often correct. However, if the tool classifies someone as high risk, this is quite often incorrect, and almost more than half of individuals targeted as high-risk are incorrectly classified (Fazel et al., 2012). False positives (defendants are predicted to re-offend, but they do not) seem to be more common than false negatives (defendants are predicted not to re-offend, but they do) (Fazel et al., 2012).

The result is that many people may be or remain incarcerated, while they do not pose a danger to society. As Fazel et al. (2012) wrote: "One implication of these findings is that, even after 30 years of development, the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence-based." This diagnosis of the current state of affairs makes it important to look for ways to improve risk assessment in forensic psychiatry and criminal justice.

Algorithms hold the promise of performing more accurate predictions of criminal behavior than classic approaches, commonly derived from various forms of regression analyses (Berk and Hyatt, 2015). They can be used to provide measures of individualized risk for future violence and help to make decisions about prevention and treatment, in order to minimize risk factors and accentuating protective ones. Risk assessment tools that incorporate machine learning are already in use in pretrial risk evaluation, sentencing, and rehabilitation (Kehl et al., 2017), and are potentially very useful in judicial decision-making, to guide "decisions regarding bail, probation/parole, court-ordered treatment, and civil commitment" (Poldrack et al., 2018).

A.I. AND NEUROIMAGING

Rapid advances in brain imaging and the growing influence of A.I. technologies in many areas of society, from social networks to health care and police force policies (Berk et al., 2018), have led to interest in the potential use of brain imaging combined with A.I. to improve risk assessment and prediction of future violent behavior.

Over the past decade, there has been a significant development of non-invasive anatomical and functional neuroimaging technologies, yielding a lot of data, and statistical machine learning methods are instrumental for analyzing vast amounts of neural data with increasing precision (Lemm et al., 2011) and modeling high-dimensional datasets (Abraham et al., 2014).

Applying statistical machine learning methods to neuroimaging data is referred to as multi-voxel pattern analysis (MVPA) (Ombao et al., 2017, pp164–169). These methods, unlike conventional univariate approaches that analyze only one location at a time, allow for the identification of spatial and temporal patterns in the data, differentiating between cognitive tasks or subject groups with higher sensitivity, jointly analyzing data from individual voxels within a region (Haynes and Rees, 2006).

Since the advent of MVPA methods, they have become a popular approach in the "neuroimaging of healthy and clinical populations; studies have shown that information present in neuroimaging data can be used to decode" – to some extent – "intentions and perceptual states, as well as discriminate between healthy and diseased brains" (Bray et al., 2009). MVPA has been applied to decode visual features like edge orientation (Kamitani and Tong, 2005), the intention to perform one task rather than another (Haynes et al., 2007), sequential stages of task preparation (Bode and Haynes, 2009), and lie detection (Davatzikos et al., 2005; Blitz, 2017, pp. 45–58). While conventional functional imaging studies compare brain activity during different experimental conditions to identify which brain regions are activated by particular tasks, application of MVPA for brain-reading uses "patterns of brain activity to perform a reverse inference and decide what subjects are looking at or thinking about" (Cox and Savoy, 2003; Bray et al., 2009).

These techniques can be considered 'brain-reading' or 'mind-reading' techniques; they combine statistical machine-learning methods with neuroimaging data to reveal information about the brain/mind. Brain-reading has often been studied in the domain of visual perception, where it aims to show how experiences are encoded in the brain. Researchers recently succeeded in training a deep neural network[1] to perform visual image reconstruction from the brain (Shen et al., 2019), decode visual content of dreams (Horikawa et al., 2013), and decode what the brain is 'seeing' by using A.I. to analyse fMRI scans from subjects watching videos (Wen et al., 2017). Despite promising findings, these methods still show many limitations that make it unlikely that a 'general mind-reading technique' will appear in the very near future. Nonetheless, the first simple applications have begun to emerge, including brain-computer-interfaces, studies on lie-detection and approaches for prediction of consumer decisions in the field of neuromarketing (Haynes, 2012, pp. 29–40).

Apart from making inferences regarding the occurrence and nature of mental states (Haynes, 2012, pp. 29–40), another field of application of MVPA techniques is classification. For example, it has been reported that it is possible to predict disease onset by distinguishing individuals within a group based on brain activity or classifying individual people into groups based on the brain data identifying patterns of brain activity or structures (Koutsouleris et al., 2012). Treatment responders can be distinguished from non-responders, by extracting patterns of activity or structural abnormalities that are predictive of abnormal cognitive development and particularly relevant for prediction of clinical outcomes from neuroimaging data (Bray et al., 2009). Some models are applied to discriminate between clinical groups such as Alzheimer Disease patients and cognitively normal elderly individuals (Klöppel et al., 2008), Parkinson's disease patients and healthy controls (Rubbert et al., 2019), schizophrenic patients and healthy controls (Kim et al., 2016), or to detect brain function disorders, such as Autism and attention deficit hyperactivity disorder (ADHD) (Heinsfeld et al., 2018; Sen et al., 2018) and to discriminate between levels of personality traits, for example psychopathy (Steele et al., 2015).

Interesting results have also been reported about prediction of addiction outcomes; machine learning classifiers were able to predict substance abuse treatment completion in a prison inmate population using event-related potentials (ERPs) (Steele et al., 2014; Fink et al., 2016) and functional network connectivity (FNC) analyses of fMRI data (Steele et al., 2018). Furthermore, it turned out to be possible to identify 'neural fingerprints' to predict cocaine abstinence during treatment using CPM, a recently developed machine learning approach (Yip et al., 2019).

A.I. NEUROPREDICTION OF RECIDIVISM

Behavioral traits can be correlated, sometimes strongly, with features of the human brain, and this raises new possibilities for predictive algorithms to be developed, allowing the prediction of dispositions of an individual. These methods are referred to as "neuroprediction," that is the use of structural or functional brain variables to predict prognoses, treatment outcomes, and behavioral forecasts (Morse, 2015). Even though at present it may sound like science fiction, with the continuing development of non-invasive neuroimaging techniques coupled with the growth in the computational power of algorithms, A.I. neuroprediction of recidivism is likely to become available in the near future.

Although there is still need to collect biomarkers of the "criminal" brain, research in the field of neurocriminology has generally focused on the analysis of structural and functional neuromarkers of personality disorders whose main characteristic consists of persistent antisocial conduct, such as ASPD (De Brito et al., 2009) and psychopathy (Umbach et al., 2015), because they appear to be the most correlated to high rates of recidivism (Coppola, 2018). Research shows that these particular clinical populations share many traits, such as behavioral disinhibition or a lack of empathy, that are supposed to have common neurobiological bases (Coppola, 2018).

For example, abnormalities in limbic and paralimbic regions have been observed in individuals with psychopathic traits (Anderson and Kiehl, 2012) and impairments related to the prefrontal cortex are associated with disinhibition, emotional lability, and impulsivity (Chow, 2000; Yang and Raine, 2009).

Still, all such neurocriminological findings, obtained using conventional methods, do not enable us at this moment to make predictions of future risk. However, incorporating neurodata in A.I. prediction models appears to open up this possibility.

[1] A neural network is "a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes." [DARPA Neural Network Study (U.S.)., United States. Air Force. Systems Command., Lincoln Laboratory. (1989). DARPA neural *network study final report.* Lexington, Mass.: The Laboratory].

A first step toward A.I. prediction models using neuroimaging data is a study conducted by Aharoni et al. (2013), who used fMRI data to predict recidivism. The authors showed that activation in the dorsal anterior cingulate cortex (dACC), a brain region associated with impulse control and error processing, during a go/no-go task appeared to be associated with rearrest. The probability that offenders with relatively low anterior cingulate activity would be rearrested was approximately double compared to an offender with high activity in this region, keeping all the other risk factors constant. Low anterior cingulate activity, therefore, might be a potential neurocognitive biomarker for persistent criminal behavior (Aharoni et al., 2013).

Recently, a study by Kiehl et al. (2018) used machine learning coupled with neuroimaging to test whether brain age could help predict rearrest. Chronological young age is considered one of the key risk factors for recidivism. Young defendants are more likely to engage in risky behavior. Kiehl proposes that brain age is a better measure to account for individual differences than chronological age. The results of his study show that a predictive model involving neural measures of brain age performed better than previous models including only psychological and behavioral measures.

Even more recently, a study by Delfin et al. (2019) shows that improvements in recidivism prediction in forensic psychiatry might be possible by incorporating neuroimaging data into A.I. risk assessment models. The authors showed that the inclusion of resting-state regional cerebral blood flow (rCBF) measurements in an extended A.I. prediction model, containing neural measurements from eight brain regions, leads to an increase in predictive performance over traditional, empirical risk factors in a long-term follow-up of forensic psychiatric patients. Interestingly, they used ‘classical’ risk assessment *combined* with neuroimaging, which showed a better prediction in a forensic psychiatric population than the classical factors alone (Delfin et al., 2019).

In sum, preliminary findings in A.I. neuroprediction studies have produced some promising results. Still, the possible use of A.I. and ‘brain-reading’ in forensic populations raises several ethical and legal concerns, and the field of criminal justice should be cautious about their future use.

It is crucial to balance the preservation of offenders’ individual rights on the one hand and the enhancement of public safety on the other.

PREDICTIVE ANALYSIS: TECHNOLOGICAL LIMITATIONS AND PITFALLS

Despite the opportunities previously discussed regarding the future possible use of A.I. neuroprediction techniques, several limitations should be considered; indeed, research about prediction tools and their successful application is still a challenging task (Poldrack et al., 2019).

This issue is well-known in the field of computational psychiatry, in which studies combining machine learning approaches and neuroimaging-based single subject prediction of brain disorders aim to classify patients with heterogeneous disorders (Arbabshirani et al., 2017; Bzdok and Meyer-Lindenberg, 2018). These studies, interestingly, reported varying degrees of accuracy (Neuhaus and Popescu, 2018), raising concerns about the methodology (Cearns et al., 2019). In fact, there is a need for best practices in predictive modeling (Poldrack et al., 2019); a problem of neuroprediction models is that, even though they can manage complex data such as brain imaging scans, they need best practices to ensure enough statistical power to test them (Varoquaux, 2018). Several issues deserve attention here.

First, application of neuroprediction techniques requires an inference from group-level to individual predictions (Hahn et al., 2017). Another challenge concerns validation of the results in a new group – different from the data set that was used to train the algorithm. The validity of prediction models is assessed by their ability to generalize; for most learning algorithms, the standard practice is to estimate the generalization performance through a process called ‘cross-validation’: the dataset is split into two sets, a training set, used to fit the model, and a test set (Hastie et al., 2009; Varoquaux, 2018), and subsets of the data are used to train and test the predictive performance of the model iteratively.

Notably, the use of cross-validation with small samples can lead to highly variable and inflated estimates of predictive accuracy (Luedtke et al., 2019; Poldrack et al., 2019). Training machine learning algorithms requires large amounts of data; using a limited sample size may cause so-called *overfitting*, in which the model fits perfectly to the specific data set used to train it, but fits poorly to new and unseen data (Hastie et al., 2009; Poldrack et al., 2019). There is still no agreement on the adequate size of the dataset (Cearns et al., 2019); Luedtke et al. (2019) recommend to perform prediction analyses with samples no smaller than several 100 observations. Acquiring many samples, however, is often difficult and costly, especially when neuroimaging data are involved (Arbabshirani et al., 2017).

ETHICAL AND LEGAL CHALLENGES

Prediction of recidivism using A.I. neuroprediction techniques evokes ethical and legal concerns, but also new possibilities. In what follows, we discuss some central ethical and legal issues.

First, we are confronted with the issue of *bias*. Since the advent of algorithmic risk assessment, a lot of reports have documented the fact that they are “dangerously” biased. The most famous case of supposed A.I. prejudice was reported by ProPublica in May 2016. COMPAS, an algorithm widely used in the US to guide sentencing by predicting the likelihood of a criminal reoffending, turned out to be racially biased against black defendants, according to ProPublica, because they were more likely than white defendants to be incorrectly classified as high risk (“false positives”)[2] (Angwin et al., 2016). More recently, COMPAS has also been depicted as a “sexist algorithm” because its

[2] The company that produced the Compass algorithm, Northpointe, claimed in a report that the accuracy in the prediction of violence for both groups of defendants was the same: around 70% of crimes were predicted correctly (see Dieterich et al., 2016, COMPAS risk scales: demonstrating accuracy equity and predictive

algorithmic outcomes seem to systemically overclassify women in higher-risk groups (Hamilton, 2019). Similarly, Predpol, an algorithm designed to predict when and where crimes will take place, already in use in several US states, in 2016 – after an analysis of the Human Rights Data Analysis Group – was found to result in police *unfairly* targeting certain neighborhoods. Officers were repeatedly sent to areas of the city with a high proportion of people from racial minorities, regardless of the effective true crime rate in those areas (Ensign et al., 2018). Furthermore, facial recognition software, increasingly used in law enforcement, represents another potential source of both race and gender bias (Raji and Buolamwini, 2019*)*. Another example concerns Amazon's 'Rekognition' software, which is used by some police departments and other organizations. In 2018, the ACLU found that it incorrectly matched members of the Congress with people who had been charged with a crime, disproportionally misidentifying African-American and Latino members of Congress as the people in mug shots[3]. A recent study evaluating the accuracy of three commercial gender classifiers showed that they performed better in classifying male subjects than female subjects, and all of them performed worst on darker-skinned females (Buolamwini and Gebru, 2018). Moreover, recent studies show that, if left unchecked, word embeddings A.I. exhibit outdated gender stereotypes, such as "*doctors*" being male and "*receptionists*" being female (Bolukbasi et al., 2016).

These findings have led to a broader debate about the *fairness* of risk assessment using A.I. (Berk et al., 2018). Although algorithmic risk assessments can be perceived as a means of overcoming human bias, they could still reflect prejudice and institutionalized bias. A.I. is trained on data – for example, criminal files – that may themselves reflect biases on the part of police officers, prosecutors, or judges. Based on these data, the algorithm then "concludes" that groups with certain traits are more dangerous than others, while in fact, this is the result of biased data. This sometimes is referred to as "bias in-bias out." The results of A.I. prediction, in other words, highly depend on the quality of the data used. One advantage of using neuroimaging data – instead of police files – might be that neuroimaging does not reflect human bias. A.I. looks for correlations between brain activity and recidivism. Therefore, A.I. neuroprediction may offer possibilities to *decrease* bias in risk assessment. However, also since neuroprediction may be incorporated in existing risk assessment tools (see the study by Delfin et al., 2019), bias will remain a problem as long as there is no solution to bias in algorithms in general.

Furthermore, we should keep in mind that risk assessment is "quintessentially discriminatory" (Binns, 2017), meaning that it is all about classifying subjects into groups of low or high-risk individuals based on group traits. Neuromarkers for recidivism will undoubtedly be more prevalent in certain groups than in others. Treating groups of people differently because of their "brain" raises difficult questions about what constitutes unjustified unequal treatment. This question, however, is not typical of A.I. neuroprediction, but is a central issue in risk assessment and fairness in general (Nadelhoffer et al., 2012, p. 95; Tonry, 2014). Classifying people into groups based on their brain scan, even if useful to prevent possible harms, could easily lead to stigmatization and discriminating effects for those considered "high risk" in other aspects of the individual's life. It could become a sort of *modern phrenology*, by discriminating between people based on what their brain looks like. While certain institutional procedures could discriminate against those considered "high risk," stigmatization could be a more social process that excludes certain individuals based on their risk profile; for instance, stigmatization may be a consequence of sex offenders' registration (Tewksbury, 2005).

A second point concerns *privacy*. The neurodata and other data used to predict recidivism can clearly also be of interest for other purposes. For instance, for insurance companies, or when screening job applicants. Who should have access to these data, and under which conditions? Should insurance companies have access to them, and if not, should they be able to request such a procedure in order to assess the risk of a particular candidate client? Clearly, in this case, data protection – and possible access – is a fundamental issue, already highly debated in algorithms used in the era of big data. Obviously, there is also a parallel with the current debate on the nature of consent and the degree of control citizens have regarding health information in biobanks. The discussion of commercialization of genetic/health information and rights of control ("biorights") are likely to intensify in the coming years (see also Caulfield and Murdoch, 2017).

A third, related point concerns the probability of a *negative 'self-fulfilling prophecy*.' This qualm comes from recent studies, showing that receiving genetic risk information can actually *influence* your behavior, physiology, and subjective experience and change your overall risk profile (Turnwald et al., 2019). Researchers from Stanford University found that when people were told of a genetic tendency for either obesity or lower exercise capacity, acquiring this information had a physiological impact on their bodies, modifying how they responded to a meal or to exercise. A persistent discovery was that perceptions of risk altered health outcomes, therefore those informed of having the high-risk gene had a worse outcome than those informed of having the protective one (Turnwald et al., 2019). Following these findings, one may wonder how the mindset of people may be affected when you inform them about their own risk information, either genetic or neural, and how this could actually alter their risk profile. This shows that providing information may also require ethical and/or legal research and regulation.

Furthermore, it is still not clear how to exactly classify and conceptualize neurodata as risk factors. For example, in a study by Kiehl et al. (2018), a measure of brain age (gray matter) is used to predict recidivism. Chronological age is often considered a static factor, but when referring to brain measures,

parity. Retrieved from www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html). The different levels of false positives among black defendants and white defendants were to be attributed, according to Northpointe, to different base rates in the prevalence of crime among black and white defendants. It is possible to have the algorithm acquire the same level of false positives over groups with a different base rate. However, this comes at the cost of reduced accuracy. There is an extensive literature on fairness in A.I. prediction, and its trade-offs (Berk et al., 2018). The text about these algorithms is partially based on Cossins (2018).

[3] https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

we should reflect on how they should be conceptualized among risk factors. For instance, given the plasticity of the brain, should we consider brain age as a dynamic or static risk variable? How do we evaluate an offender if, for example, brain age and normal age differs, and how would this modify his/her neuroprediction profile? If we consider neurodata as dynamic factors, and, as such, available to be modified through interventions, we could talk, instead of in terms of a pure "prediction," in terms of targets for treatment and other intervention types. Used in this way, neuroprediction could help to prevent crime through more individualized correctional and socio-rehabilitative measures, and could also enable offenders to return to the community sooner. As in "personalized medicine" – a therapeutic approach in which an individual's genetic and epigenetic information is used to tailor drug therapy or preventive care[4] – neuroprediction could help to target interventions to the individual's "needs."

There is another effect of the emphasis on prediction that is relevant here. Currently, A.I. is used in the criminal justice system, mainly to predict recidivism. A.I. risk assessment typically does not offer a causal model of crime and therefore, is not designed to show opportunities to intervene and to mitigate risk (Berk, 2019, pp. 17–18). Barabas et al. (2018) conclude: "when risk assessments are used primarily as a predictive technology, they fuel harmful trends toward mass incarceration and growing inequality in the justice system."

We should acknowledge that A.I. neuroprediction in the first place merely establishes correlations between brain images and the risk of recidivism. However, if it is indeed possible to develop interventions based on neurodata, this might offer offenders an opportunity to avoid incarceration (Nadelhoffer et al., 2012, pp. 85–86). This could be possible because, different from historical data and other risk variables, like a person's demographic characteristics such as ethnicity, age, and gender, that cannot be changed, neurodata hold the potential to become targets for new rehabilitative interventions and prevention programs, aiming to reduce exposure to risk factors for psychopathic traits and preventing at-risk individuals from engaging in criminal behavior later in life (Ling and Raine, 2018).

This is particularly important since the prison environment may have negative effects on neurocognitive functioning. In fact, studies found that incarceration might lead to reduced self-control (Meijers et al., 2018). Still, the possibility of intervention also entails its own ethical and legal issues: for an offender, it may be hard to choose between a deprivation of liberty and undergoing (possibly somewhat invasive) treatment, especially in light of the right to refuse medical treatment (Meynen, 2018). However, this again is not a problem that is typical of interventions based on "A.I. neuroprediction."

A fourth, and related, issue concerns *consent and coercion*; if and when these techniques will be fully developed and are ready to be used, there may be a possibility of performing cognitive liberty violations forcing people to undergo scans without consent for sentencing or punitive purposes (Ligthart, 2019; Meynen, 2019). Coercion, both technical and ethical or legal, not only relates to the force used, because not all the imaging techniques allow for this, but also to their use within the context of a threat or an offer that cannot be refused (Meynen, 2017). One way to counter this issue is to strictly regulate informed consent for neuroprediction tests.

Fifth, we should take into account something called the "seductive allure" that neuroimaging exerts on courts. Juries and judges apparently tend to overestimate the accuracy of neuroscientific evidence, and, although neuroimaging aims to reduce uncertainty and to increase the objectivity in forensic settings, the use of neuroimaging in courts is at risk of being misleading, due to cognitive biases in the evaluation of evidence (Scarpazza et al., 2018). Introducing neuroprediction could therefore lead to some overreliance on neurodata.

Furthermore, machine learning algorithms are considered to be '*black-boxes of decision-making*'; the way in which they perform decisions is not fully comprehensible to stakeholders, and not even to expert data scientists (London, 2019; Pedreschi et al., 2019). In addition, we have to be cautious about what is called the "the control problem"; i.e., the tendency of human operators to become complacent with machines, devolving responsibility and becoming over-reliant on the outputs of autonomous systems, even when they are biased (Pedreschi et al., 2019). In order to avoid overreliance, it seems important for A.I. systems to be transparent: it should be possible to explain to judges and a jury how they produce their results (Gunning and Aha, 2019), and stakeholders should be capable to appropriately trust and manage these tools, reasoning on how a specific output is given and on the basis of what rationale (Pedreschi et al., 2019). Even if this is actually complicated by the fact that most risk assessment algorithms are proprietary, it seems important for society that A.I. algorithms can be made intelligible, in order to be accountable for their decisions (Weld and Bansal, 2019).

Of note, legal systems may have criteria for the admissibility of scientific evidence in the courtroom. For instance, in the US legal context *Daubert* and *Frye* are used as standards. As we do not focus on specific legal systems, we will not go into this in more detail, but clearly such legal criteria would be relevant for courtroom use of new technologies (Shats et al., 2016).

Moreover, it is important to make a decision about the required accuracy of these technologies. Current risk assessment tools often have an AUC of about 0.70 (Douglas et al., 2017); is that enough for such algorithms, or should the threshold be higher, like 0.80 or 0.90? These are normative choices that have to be made before deciding to allow the use of this kind of technology to prevent crime.

Additionally, we need to consider the lack, at present, of a 'true' prediction model; a limitation of the papers previously discussed is that, instead of talking about 'pure' prediction, they can be classified as *postdiction* studies; postdiction generally relates to retrospectively making an assertion or deduction about an event based on information available after the event (Yamada et al., 2015) but, as applied to the context of statistical models, the distinction between prediction and postdiction is about whether the assessment of the model's success involves the same data as were used to build the model or new data not used

[4] https://www.nature.com/subjects/personalized-medicine

in model construction (Gauch and Zobel, 1988; Hastie et al., 2009). Research suggests that models for predictive applications, such as biomarkers, require larger sample sizes than standard statistical approaches (Varoquaux, 2018). Furthermore, in the studies discussed before, data about neuromarkers of recidivism have been collected after the commission of crimes, so we cannot establish when brain differences observed developed (Cope et al., 2014). A future challenge is to develop a true prediction model, able to identify those at the highest risk for committing crimes, and research in neuroimaging coupled with A.I. may be the key in developing such model.

Finally, there appears to be a more remote problem, looming on the horizon. Suppose that these A.I. algorithms – either with or without brain imaging – become really good predictors, wouldn't that introduce a form of determinism we have not witnessed before? The A.I. system may be considered to have some "divine" foreknowledge about what will happen, which may have negative effects on the freedom people experience and exert. A belief in free will seems to have positive effects (Crescioni et al., 2016; Feldman et al., 2016).

Still, the more pressing concern nowadays is that we are not quite good at predicting risk – even with A.I. – and that we nonetheless often apply sanctions based on the supposed dangerousness of the offender. If A.I. becomes more accurate with the help of neuroimaging, it could reduce the number of persons incorrectly classified as high risk and can therefore reduce sanctions that in fact are not legitimate, helping to interrupt the so-called "cycles of crime" (Barabas et al., 2018).

CONCLUSION

There is still a way to go before combined neuroscience and AI-based violence risk assessment tools can be implemented in the criminal justice system. Still, A.I. is already being used in criminal justice systems. Because of the far-reaching consequences of these type of technologies – and also given some rapid developments in recent years – it is important to consider ethical and legal concerns. Besides discussing technological limitations and pitfalls of predictive analysis, we identified six key issues deserving attention: dealing with bias, privacy, the possibility of a 'self-fulfilling prophecy,' coercion and consent, the allure of neuroimaging data and the need for A.I. systems to be explainable. Finally, we pointed to the more remote issue of how highly accurate predictions might introduce a form of determinism we have not witnessed before – but this is still far away.

Still, we would like to emphasize that accurate risk prediction is extremely valuable for both safety and justice reasons. Therefore, in principle, we argue that technologies that may be helpful in this respect should at least be explored, and if ready, used in criminal justice and forensic psychiatry. In addition, neuroprediction and A.I. bring their own, in a way new, ethical and legal challenges, and we will have to deal with them – preferably before the technologies are used. More specifically, we have to find solutions to prevent systems from reflecting our own human biases in order to enable them to provide objective and trustworthy data.

Therefore, we argue that the use of AI-based systems in criminal justice and forensic psychiatry should be subjected to substantial regulation to protect citizens from system errors or misuse. On such basis, we highlight the importance of accurate harms/benefits analyses not only when these technologies will be fully available, but also while they are being researched and developed.

AUTHOR CONTRIBUTIONS

LT, GM, and SF conceived the content of the manuscript and wrote and revised the manuscript. LT drafted the manuscript. JB and ET wrote and revised the manuscript. All authors read and approved the final manuscript.

REFERENCES

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* 8:14. doi: 10.3389/fninf.2014.00014

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W., Gazzaniga, M. S., et al. (2013). Neuroprediction of future rearrest. Proceedings of the national academy of sciences of the united states of america. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6223–6228. doi: 10.1073/pnas.1219302110

Anderson, N. E., and Kiehl, K. A. (2012). The psychopath magnetized: insights from brain imaging. *Trends Cogn. Sci.* 16, 52–60. doi: 10.1016/j.tics.2011.11.008

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias*. New York, NY: ProPublica.

Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage* 145(Pt B), 137–165. doi: 10.1016/j.neuroimage.2016.02.079

Barabas, C., Virza, M., Dinakar, K., Ito, J., and Zittrain, J. (2018). "Interventions over predictions: reframing the ethical debate for actuarial risk assessment," in *Proceedings of FAT conference (FAT 2018). ACM*, New York, NY, 62–76.

Berk, R. (2019). *Machine Learning Risk Assessments in Criminal Justice Settings*. Berlin: Springer.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: the state of the art. *arXiv.org* [Preprint], doi: 10.1177/0049124118782533

Berk, R., and Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Fed. Sentenc. Rep.* 27, 222–228. doi: 10.1525/fsr.2015.27.4.222

Binns, R. (2017). Fairness in machine learning: lessons from political. *Philos. Proc. Mach. Learn. Res.* 81, 1–11.

Blitz, M. J. (2017). "Lie detection, mind reading, and brain reading. in: searching minds by scanning brains," in *Palgrave Studies in Law, Neuroscience, and Human Behavior*, (Cham: Palgrave Macmillan), 45–58. doi: 10.1007/978-3-319-50004-1_3

Bode, S., and Haynes, J.-D. (2009). Decoding sequential stages of task preparation in the human brain. *Neuroimage* 45, 606–613. doi: 10.1016/j.neuroimage.2008.11.031

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inform. Proc. Syst.* 20, 4349–4357.

Bray, S., Chang, C., and Hoeft, F. (2009). Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. *Front. Hum. Neurosci.* 3:32. doi: 10.3389/neuro.09.032.2009

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Proceeding of F.A.T*, New York, NY.

Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry* 3, 223–230.

Caulfield, T., and Murdoch, B. (2017). Genes, cells, and biobanks: yes, there's still a consent problem. *PLoS Biol.* 15:e2002654. doi: 10.1371/journal.pbio.2002654

Cearns, M., Hahn, T., and Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* 9:271. doi: 10.1038/s41398-019-0607-2

Chow, T. W. (2000). Personality in frontal lobe disorders. *Curr. Psychiatry Rep.* 2, 446–451. doi: 10.1007/s11920-0000031-5

Conroy, M. A., and Murrie, D. C. (2007). *Forensic Assessment Of Violence Risk: A Guide For Risk Assessment And Risk Management*. Hoboken, NJ: John Wiley & Sons Inc, doi: 10.1002/9781118269671

Cope, L. M., Ermer, E., Gaudet, L. M., Steele, V. R., Eckhardt, A. L., Arbabshirani, M. R., et al. (2014). Abnormal brain structure in youth who commit homicide. *Neuro. Clin.* 4, 800–807. doi: 10.1016/j.nicl.2014.05.002

Coppola, F. (2018). Mapping the brain to predict antisocial behaviour: new frontiers in neurocriminology, 'new'challenges for criminal justice. *U.C.L. J. Jurisprud. Spec.* 1, 106–110.

Cossins, D. (2018). Discriminating algorithms: 5 times AI showed prejudice. *New Scientist* (accessed January 10, 2019).

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/s1053-8119(03)00049-1

Crescioni, A. W., Baumeister, R. F., Ainsworth, S. E., Ent, M., and Lambert, N. M. (2016). Subjective correlates and consequences of belief in free will. *Philos. Psychol.* 29, 41–63.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., et al. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 28, 663–668. doi: 10.1016/j.neuroimage.2005.08.009

De Brito, S. A., Mechelli, A., Wilke, M., Laurens, K. R., Bartoli, A. J., Barker, G. J., et al. (2009). Size matters: increased grey matter in boys with conduct problems and callous-unemotional traits. *Brain* 132(Pt 4), 843–852. doi: 10.1093/brain/awp011

Delfin, C., Krona, H., Andine', P., Ryding, E., Wallinius, M., and Hofvander, B. (2019). Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: incremental effects of neuroimaging data. *PLoS One* 14:e0217127. doi: 10.1371/journal.pone.0217127

Dieterich, W., Mendoza, C., and Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Northpointe Inc.

Douglas, K. S., and Kropp, P. K. (2002). A prevention-based paradigm for violence risk assessment: clinical and research applications. *Crim. Just. Behav.* 29, 617–658. doi: 10.1177/009385402236735

Douglas, T., Pugh, J., Singh, I., Savulescu, J., and Fazel, S. (2017). Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. *Eur. Psychiatry* 42, 134–137. doi: 10.1016/j.eurpsy.2016.12.009

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). "Runaway feedback loops in predictive policing," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Berlin.

Fazel, S., Singh, J. P., Doll, H., and Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *B.M.J.* 345:e4692. doi: 10.1136/bmj.e4692

Feldman, G., Chandrashekar, S. P., and Wong, K. F. E. (2016). The freedom to excel: belief in free will predicts better academic performance. *Pers. Indiv. Differ.* 90, 377–383.

Fink, B. C., Steele, V. R., Maurer, M. J., Fede, S. J., Calhoun, V. D., and Kiehl, K. A. (2016). Brain potentials predict substance abuse treatment completion in a prison sample. *Brain Behav.* 6:501. doi: 10.1002/brb3.501

Gauch, H. G., and Zobel, R. W. (1988). Predictive and postdictive success of statistical analyses of yield trials. *Theoret. Appl. Genetics* 76, 1–10. doi: 10.1007/BF00288824

Gendreau, P., Little, T., and Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: what works. *Criminology* 34, 575–608. doi: 10.1111/j.1745-9125.1996.tb01220.x

Gunning, D., and Aha, D. (2019). DARPA's explainable artificial intelligence (X.A.I.) Program. *A.I. Magaz.* 40, 44–58. doi: 10.1609/aimag.v40i2.2850

Hahn, T., Nierenberg, A., and Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol. Psychiatry* 22, 37–43. doi: 10.1038/mp.2016.201

Hamilton, M. (2019). The sexist algorithm. *Behav. Sci. Law* 37, 145–157. doi: 10.1002/bsl.2406

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: conceptual and methodological issues. *Legal Criminol. Psychol.* 3, 121–137. doi: 10.1111/j.2044-8333.1998.tb00354.x

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Berlin: Springer, doi: 10.1007/978-0-387-84858-7

Haynes, J.-D. (2012). "Brain reading," in *I Know What You're Thinking: Brain imaging and Mental Privacy*, eds D. Sarah, G. Rees, and J. L. Sarah, (Oxford: Oxford University Press), doi: 10.1093/acprof:oso/9780199596492.003.0003

Haynes, J. D., and Rees, G. (2006). Neuroimaging: decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534.

Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr. Biol.* 17, 323–328.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuro. Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science* 340, 639–642. doi: 10.1126/science.1234330

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685. doi: 10.1038/nn1444

Kehl, D., Guo, P., and Kessler, S. (2017). *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities*. Available online at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041 (accessed March 1, 2019).

Kiehl, K. A., Anderson, N. E., Aharoni, E., Maurer, J. M., Harenski, K. A., Rao, V., et al. (2018). Age of gray matters: neuroprediction of recidivism. *Neuroimage* 19, 813–823. doi: 10.1016/j.nicl.2018.05.036

Kim, J., Calhoun, V. D., Shim, E., and Lee, J. H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage* 124(Pt A)), 127–146. doi: 10.1016/j.neuroimage.2015.05.018

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* 131(Pt 3), 681–689. doi: 10.1093/brain/awm319

Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Möller, H. J., and Riecher-Rössler, A. (2012). Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophr. Bull.* 38, 1234–1246. doi: 10.1093/schbul/sbr145

Lemm, S., Benjamin, B., Thorsten, D., and Mueller, K. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Ligthart, S. (2019). "Coercive neuroimaging technologies in criminal law in Europe: exploring the implications for the prohibition of ill-treatment (article 3 ECHR)," in *Regulating New Technologies In Uncertain Times Information Technology And Law Series*, ed. L. Reins, (Berlin: Springer), 83–102. doi: 10.1007/978-94-6265-279-8-6

Ling, S., and Raine, A. (2018). The neuroscience of psychopathy and forensic implications. *Psychol. Crim. Law* 24, 296–312. doi: 10.1080/1068316X.2017.1419243

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hast. Center Rep.* 49, 15–21. doi: 10.1002/hast.973

Luedtke, A., Sadikova, E., and Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin. Psychol. Sci.* 7, 445–461. doi: 10.1177/2167702618815466

Meijers, J., Harte, J. M., Meynen, G., Cuijpers, P., and Scherder, E. J. A. (2018). Reduced self-control after 3 months of imprisonment. A pilot study. *Front. Psychol.* 9:69. doi: 10.3389/fpsyg.2018.00069

Meynen, G. (2017). Brain-based mind reading in forensic psychiatry: exploring possibilities and perils. *J. Law Biosci.* 4, 311–329. doi: 10.1093/jlb/lsx006

Meynen, G. (2018). Forensic psychiatry and neurolaw: description, developments and debates. *Int. J. Law Psychiatry.* 65:101345. doi: 10.1016/j.ijlp.2018.04.005

Meynen, G. (2019). Ethical issues to consider before introducing neurotechnological thought apprehension in psychiatry. *AJOB Neurosci.* 10, 5–14. doi: 10.1080/21507740.2019.1595772

Monahan, J., and Skeem, J. L. (2015). Risk Assessment in Criminal Sentencing (September 17, 2015). Annual Review of Clinical Psychology, Forthcoming; Virginia Public Law and Legal Theory Research Paper, No. 53. Available online at SSRN: https://ssrn.com/abstract=2662082 (accessed January 10, 2019).

Morse, S. J. (2015). *Neuroprediction: New Technology, Old Problems. Faculty Scholarship at Penn Law.1619*. Available online at: https://scholarship.law.upenn.edu/faculty_scholarship/1619 (accessed January 10, 2019).

Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K. A., Mansfield, A., Sinnott-Armstrong, W., et al. (2012). Neuroprediction, violence, and the law: setting the stage. *Neuroethics* 5, 67–99. doi: 10.1007/s12152-010-9095-z

Neuhaus, A. H., and Popescu, F. C. (2018). Sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses. *Biol. Psychiatry* 84:e0081-2.

Ombao, H., Lindquist, M., Thompson, W., and Aston, J. (2017). *Handbook of Neuroimaging Data Analysis*. New York: Chapman and Hall/CRC.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. (2019). Meaningful explanations of black box ai decision systems. *Proc. AAAI Conf. Artif. Intellig.* 33, 9780–9784.

Poldrack, R. A., Huckins, G., and Varoquaux, G. (2019). Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry.* 27:2019. doi: 10.1001/jamapsychiatry.2019.3671

Poldrack, R. A., Monahan, J., Imrey, P. B., Reyna, V., Raichle, M. E., Faigman, D., et al. (2018). Predicting violent behavior: what can neuroscience add? *Trends Cogn. Sci.* 22, 111–123. doi: 10.1016/j.tics.2017.11.003

Raji, I., and Buolamwini, J. (2019). "Actionable auditing: investigating the impact of publicly naming biased performance results of commercial a.i. products," in *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society*, New York, NL.

Rubbert, C., Mathys, C., Jockwitz, C., Hartmann, C. J., Eickhoff, S. B., Hoffstaedter, F., et al. (2019). Machine-learning identifies Parkinson's disease patients based on resting-state between-network functional connectivity. *Br. J. Radiol.* 2019:20180886. doi: 10.1259/bjr.20180886

Scarpazza, C., Ferracuti, S., Miolla, A., and Sartori, G. (2018). The charm of structural neuroimaging in insanity evaluations: guidelines to avoid misinterpretation of the findings. *Transl Psychiatry.* 8:227. doi: 10.1038/s41398-018-0274-8

Sen, B., Borle, N. C., Greiner, R., and Brown, M. (2018). A general prediction model for the detection of ADHD and Autism using structural and functional M.R.*I. PloS one* 13:e0194856. doi: 10.1371/journal.pone.0194856

Shats, K., Brindley, T., and Giordano, J. (2016). Don't ask a neuroscientist about phases of the moon: applying appropriate evidence law to the use of neuroscience in the courtroom. *Cambridge Quarterly of Healthcare Ethics* 25, 712–725. doi: 10.1017/S0963180116000438

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS. Comput. Biol.* 15:e1006633. doi: 10.1371/journal.pcbi.1006633

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., and Dean, K. (2014). International perspectives on the practical application of violence risk assessment: a global survey of 44 countries. *Int. J. Foren. Ment. Health.* 13, 193–206. doi: 10.1080/14999013.2014.922141

Steele, V. R., Fink, B. C., Maurer, J. M., Arbabshirani, M. R., Wilber, C. H., Jaffe, A. J., et al. (2014). Brain potentials measured during a Go/NoGo task predict completion of substance abuse treatment. *Biol. Psychiatry* 76, 75–83. doi: 10.1016/j.biopsych.2013.09.030

Steele, V. R., Maurer, J. M., Arbabshirani, M. R., Claus, E. D., Fink, B. C., Rao, V., et al. (2018). Machine learning of functional magnetic resonance imaging network connectivity predicts substance abuse treatment completion. *Biol. Psychiatry* 3, 141–149. doi: 10.1016/j.bpsc.2017.07.003

Steele, V. R., Rao, V., Calhoun, V. D., and Kiehl, K. A. (2015). Machine learning of structural magnetic resonance imaging predicts psychopathic traits in adolescent offenders. *Neuroimage* 145(Pt B), 265–273. doi: 10.1016/j.neuroimage.2015.12.013

Tewksbury, R. (2005). Collateral consequences of sex offender registration. *J. Contemp. Crim. Just.* 21, 67–81. doi: 10.1177/1043986204271704

Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Fed. Senten. Rep.* 26, 167–176.

Turnwald, B. P., Goyer, J. P., Boles, D. Z., Silder, A., Delp, S. L., and Crum, A. J. (2019). Learning one's genetic risk changes physiology independent of actual genetic risk. *Nat. Hum. Behav.* 3, 48–56.

Umbach, R., Berryessa, C., and Raine, A. (2015). Brain imaging research on psychopathy: implications for punishment, prediction, and treatment in youth and adults. *J. Crim. Just.* 43, 295–306.

Varoquaux, G. (2018). Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 180(Pt A), 68. doi: 10.1016/j.neuroimage.2017.06.061

Weld, S. D., and Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 70–79. doi: 10.1145/3282486

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 28, 4136–4160. doi: 10.1093/cercor/bhx268

Yamada, Y., Kawabe, T., and Miyazaki, M. (2015). Awareness shaping or shaped by prediction and postdiction: editorial. *Front. Psychol.* 6:166. doi: 10.3389/fpsyg.2015.00166

Yang, Y., and Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Res.* 174, 81–88. doi: 10.1016/j.pscychresns.2009.03.012

Yip, S. W., Scheinost, D., Potenza, M. N., and Carroll, K. M. (2019). Connectome-based prediction of cocaine abstinence. *Am. J. Psychiatry* 176, 156–164. doi: 10.1176/appi.ajp.2018.17101147

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.