

UNITING MACHINE INTELLIGENCE, BRAIN AND BEHAVIOURAL SCIENCES TO ASSIST CRIMINAL JUSTICE

BY OLIVER Y. CHÉN

Faculty of Social Sciences and Law, University of Bristol
olivery.chen@bristol.ac.uk

This is a working paper.

For comments, criticisms, or literature/case suggestions, please email me.
I will do my best to revise based on your suggestions and criticisms.

I discuss here three important roles where machine intelligence, brain and behaviour studies together may facilitate criminal law. First, brain imaging analysis and predictive modelling using brain and behaviour data enable mental illness, insanity, and behaviour examination during legal investigations. Second, psychological, psychiatric, and behavioural studies supported by machine learning algorithms may help detect lies, biases, and visits to crime scenes. Third, brain decoding is beginning to uncover one's thoughts and intentions based on functional brain imaging data. Having dispensed with achievements and promises, I examine concerns regarding the accuracy, reliability, and explainability of the brain- and behaviour-based assessments in criminal law, as well as questions regarding data possession, security, privacy, and ethics. Taken together, brain and behaviour decoding in legal exploration and decision-making at present is promising but primitive. The derived evidence is limited and should not be used to generate definitive conclusions, although it can be potentially used in addition, or parallel, to existing evidence. Finally, I suggest that there needs to be (more precise) definitions and regulations regarding when and when not brain and behaviour data can be used in a predictive manner in legal cases.

*I thank Martin F., Václav Janeček, Marcus R. Munafò, Guy Nagels, Huy Phan, Sarah Rosanowski, Xiaojun Wang, and Bangdong Zhi for comments and criticisms on earlier versions of the paper. I thank a criminal judge (English law) for helpful discussion and suggestions.

Keywords and phrases: Machine learning, AI, brain science, behavioural science, predictive modeling, criminal justice, law.

I am a brain, Watson. The rest of me is a mere appendix. Therefore, it is the brain I must consider.

The Adventure of the Mazarin Stone

1. Introduction

All human laws are products of the brain. On the one hand, the laws are designed, defended, and delivered by involving functions of the specialised brain areas. On the other hand, understanding the functions and dysfunctions of the brain may assist jurors and judges in evaluating one's thoughts, intentions, and actions and helping them to classify wrongdoings due to irregular brain activities, poor judgments, and criminal dispositions (see **Figure 1**).

Recent development in machine intelligence, brain and behavioural sciences, and imaging technology have shown promises to use brain and behaviour signals to assess and predict various mental and cognitive faculties, such as intention ([Haynes et al., 2007](#)), mental states ([Reinen et al., 2018](#)), psychological illness ([Cao et al., 2018](#)), cognition ([Chén et al., 2019](#)), as well as their behaviour outputs, such as telling lies ([Farah et al., 2014](#)). The neurobehavioural and neuropsychological findings and their supporting technological frameworks, if proven broadly effective and reliable, may contribute to the jurisprudence in several ways: to separate false statements from facts, to exculpate those suffering from mental illness or insanity, to distinguish innocent from guilty, and to determine the degrees of fines, liabilities, or sentences.

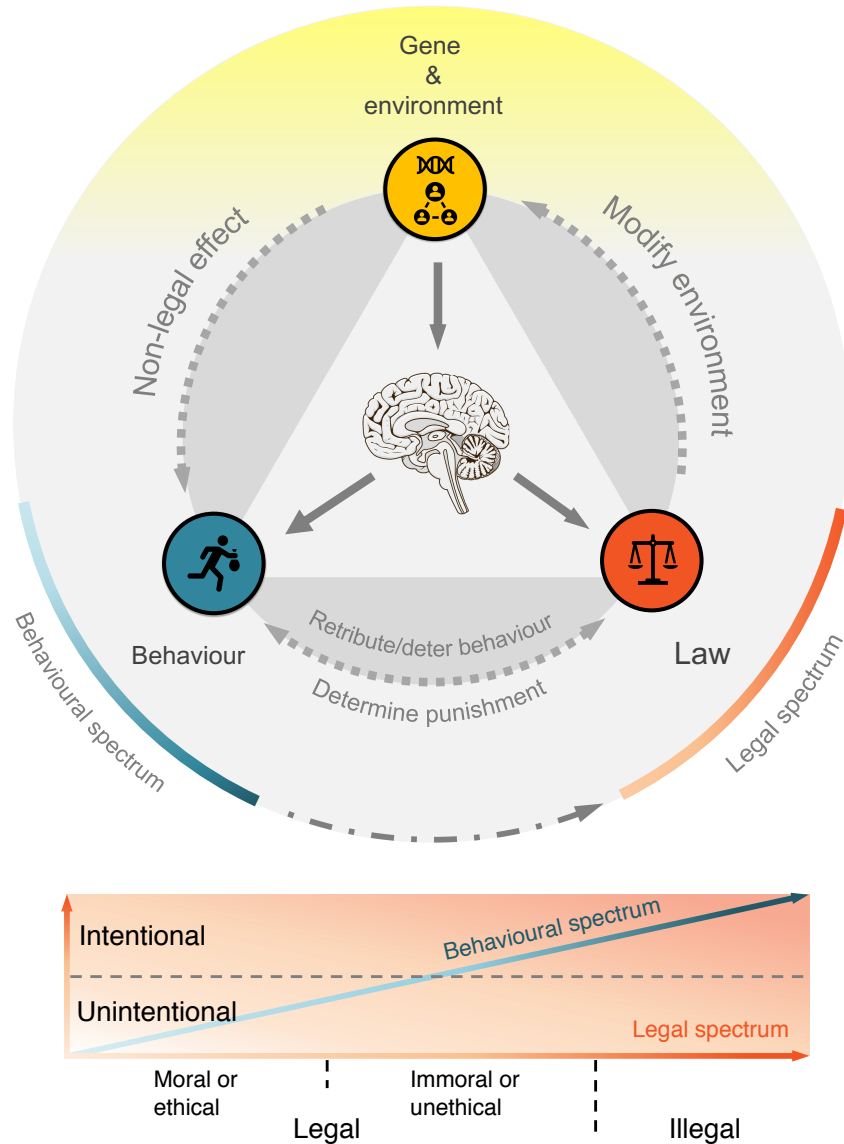
In parallel, we have seen increasing usage of AI in the criminal justice system, such as “the world’s first robot lawyer”¹, the consideration of AI-based legal solutions in Estonia², robot mediators in Canada, and AI judges in Chinese courts ([Chandran, 2022](#)). Such innovations and adaptations, however, have received mixed views. Noticeably, a recent case in Sabah, Malaysia, where a man was sentenced via the assistance of an artificial intelligence tool, has stirred legal outrage ([Chandran, 2022](#)). The supporters argue that “AI-based systems make sentencing more consistent and can clear case backlogs quickly and cheaply, helping all parties in legal proceedings to avoid lengthy, expensive and stressful litigation”. Yet the critics maintain that “it is unconstitutional”, increases bias, and that, compared to human judges, AI does not consider mitigating factors or use discretion. In the middle ground are opinions suggesting that some decisions might properly be handed over to the machines, but regulations (in the sense of public control ([Chesterman, 2021a](#))) and transparency as well as explainability ([Chesterman, 2021b](#)) are needed.

The thesis of uniting machine intelligence, brain and behaviour science to assist criminal justice, therefore, must confront three general challenges: technological difficulties, legal obstacles, and ethical concerns. For example, technologically, how reliable are predictive algorithms developed in labs when applied to real-world legal practices? Legally, even for reliable models, how should the legal profession adopt them to gather and deliver evidence during discussions and investigations? Compounding the technological and legal challenges, how can one prevent brain- and behaviour-based predictions from providing undue exculpatory evidence? Ethically, who owns our brain data; who determines whether one’s brain data

¹DoNotPay (<https://www.donotpay.com>).

²The Estonia Ministry of Justice clarifies that “Estonia does not develop AI judge[s]”; rather, “Ministry of Justice is looking for opportunities for optimization and automatization of court’s procedural steps in every types of procedures” (<https://www.just.ee/en/news/estonia-does-not-develop-ai-judge>).

can be recorded and analysed; and who judges whether the results can be used in a justice system?



Top: At the tips of the **triangle** are genetic and environmental factors, human behaviour, and law. The brain's structures and functioning are determined by genetic and environmental factors and their interactions (top arrow). The brain dictates human behaviour (left arrow) and helps design and modify the law (right arrow). In the **inner circle**, genetic and environmental factors, along with the brain, affect human behaviour (left dashed arrow). (Criminal) behaviour determines reasonable punishment one receives from the law, and in turn, the law retributes criminal human behaviour and deters future crimes (bottom dashed arrow). Law governs and changes the environment in which one lives (right dashed arrow). On the **outer circle** are the behaviour spectrum and their corresponding legal spectrum. **Bottom:** Depending on the behaviour and intention, one receives different levels of legal judgment or punishment.

Fig 1: The relationship between the brain, behaviour, and law.

Inspired by these questions and concerns, here I present the promises, challenges, potential solutions and hopeful future directions of the brain- and behaviour-based assessment, prediction, and decision-making in jurisprudence. First, I argue that machine-assisted forensic science (see definition below) may facilitate legal investigations in three directions: (1) brain data-based mental faculty and behaviour prediction; (2) lie, intention, and crime scene detection; (3) general brain decoding. Next, I discuss accuracy, privacy, and ethical challenges when employing brain- and behaviour-based decision-making in legal practices. I argue that evidence derived from neural and behavioural analyses can presently provide supportive but not conclusive information. Finally, I argue that there needs to be clear definitions and regulations regarding when and when not brain and behaviour data can be used in a predictive manner in legal cases.

Remark 1. To facilitate discussion, throughout I consider brain and behaviour data broadly. For example, I define brain data as measurements of brain signals obtained either invasively or noninvasively (such as the action potentials, BOLD fMRI³ or electroencephalogram (EEG) recordings)⁴. I define behaviour data as the general outputs of the brain that may be useful in courts, such as measurements and recordings of language and hand movement. I will also consider intermediate, intangible brain concepts that mediate⁵ brain activities and actions, such as thought and intention. They are the outcome of the brain and simultaneously the potential cause of actions.

Remark 2. Many points I discuss in this paper may be primitive or controversial. I present them regardless for two reasons. First, I outline the possibilities of using AI/machine learning and brain/behaviour research to facilitate law (and *vice versa*) and provide examples, concepts, and arguments to illustrate them. Some examples and arguments may be nascent or unusual to the norm; I do not intend to and cannot settle the linkages between AI/machine learning and the brain and behaviour and how they may evolve or advance the law in the future. A presentation of possibilities may help the readers conceptually, and I feel it is perhaps useful to expose these points as early as we can and let future research and practice contest, verify, and perhaps settle them. Second, I hope my presentations, potentially still in their infancy and contentious, may sprawl further discussions, either to modify my views or to expand my arguments.

2. On predicting brain injuries, mental illnesses, and human behaviour

Brain and behaviour studies offer promises to advance **machine-assisted forensic science**. Here, by machine-assisted forensic science, I mean the practice of using AI and machine learning algorithms to derive evidence from biological data to support decision-making in legal investigations. Our consideration of the biological data goes beyond fingerprints and DNA, which are traditionally used in legal investigations; rather, they extend to include data recorded from the brain and human behaviour through brain imaging, sensors, or wearable computers.

The disruption of the functioning and structure of the human brain may result in unusual behaviours. For example, the prefrontal cortex (PFC) is involved in executive function, cognition, control of impulsive behaviour, emotional regulation, judgment (and moral reasoning),

³Blood-oxygen-level-dependent (BOLD) functional magnetic resonance imaging (fMRI).

⁴For brain decoding, it is at present perhaps more suitable and practical to use noninvasive measurements.

⁵Such mediation pathways can be further classified hierarchically. For example, brain data → thoughts → emotions → actions.

organisation, planning, and decision-making (Stuss and Benson, 1984). When one's PFC is damaged, s/he may lose, in part or entirely, morality and propriety (Sapolsky, 2004) and present behavioural disinhibition and impaired intellectual faculty (Reber and Tranel, 2019). The studies on Phineas Gage suggested the link between frontal lobe damage and radical behaviour changes. Since then, the PFC lesions and damage have been shown to be associated with violent and criminal behaviour (Brower and Price, 2001). Besides the PFC, impairment of the amygdala may also be related to irregular emotion, decision-making, and social judgment (Anderson and Phelps, 2001, Gupta et al., 2011, Phelps, 2002).

With recent development in predictive modelling, one can now link a specific brain region (e.g., the PFC) to an **outcome of legal interest**. Here, I define a multivariate, potentially high-dimensional input (x) and an outcome of legal interest (y) as follows.

Let $x = (x_1, x_2, \dots, x_p)$ be a p -dimensional feature variable consisting of p features (e.g., brain activities from p brain areas). One could extend x to behaviour features, or features combining both behaviour and neural data.

I define an outcome of legal interest (y) as a label that shows a particular characteristic of an individual under legal investigation. The outcomes of legal interest can be, in general, grouped into one of the three types: categorical outcomes, continuous values, and longitudinal estimates (see **Figure 2** (d) for more details).

The first group of outcomes of legal interest is **categorical outcomes**. Examples of categorical outcomes of legal interest are: whether someone committed a crime intentionally or not, whether someone is guilty or not, and whether one's actions are punishable or not? Because the outcomes can take one of the defined categories (e.g., intentional vs. unintentional), they are categorical. Related to categorical outcomes of legal interest within criminal law are three types of *mens rea*: intention, recklessness, and negligence⁶. Using machine learning algorithms, one can map individual brain and behaviour data onto individual categorical outcomes to classify subjects into different groups (say, whether the individuals fall into the group with an intention to kill); a trained model can then be used to predict intentions in new subjects given their brain and behaviour data.

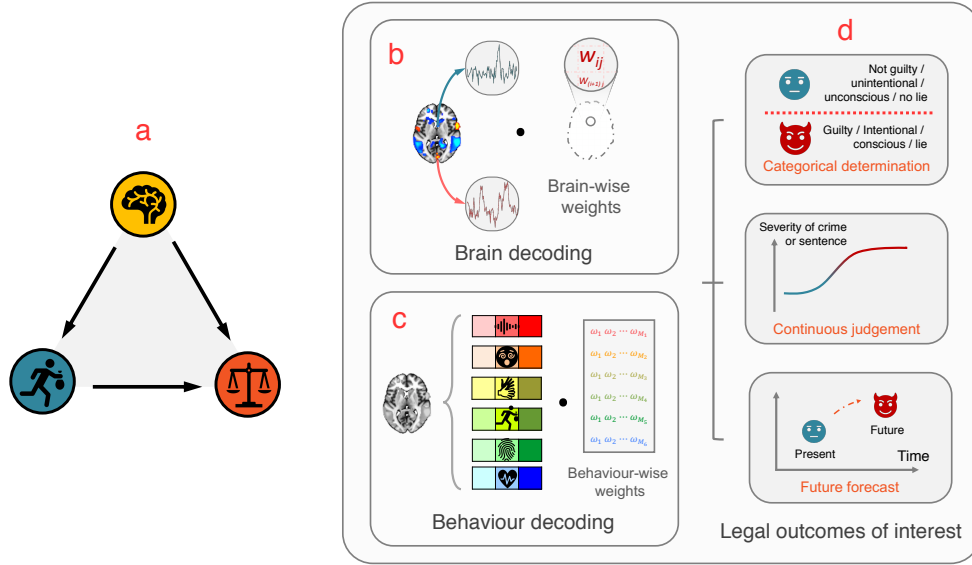
If someone intends to kill another person and kills him/her, this is viewed more harshly than if someone is reckless as to whether an act they are doing might kill someone and then it (*i.e.*, the action) kills them. To assess the severity of the outcome, one needs perhaps more refined outcomes of legal interest, that is, **continuous outcomes**. They are continuous because the outcomes can potentially take any values within a range (rather than discrete values). For example, given the brain and behaviour data, how severe (say, between 0 and 100) is one's action; how likely (say, between 0 and 1) will they yield criminal actions?

The third group is **longitudinal monitoring and forecasting**. For example, following the brain and behaviour data over time, what is the likelihood of someone committing a crime at any given point in time (monitoring), and in the future (forecasting)? Whereas the first two types of outcomes of legal interest are mostly trained and tested on observed data, the forecasting aspect deals with future data that are not as-of-yet observed. One, therefore, has to be extremely careful about predicting a person's future intentions and actions⁷ (see **Section 5**).

⁶*Intention*: A defendant will be found to have intended a consequence if they desire the consequence to follow their actions). *Recklessness*: A defendant was aware that a risk existed and went on unreasonably to take that risk). *Negligence*: A defendant does not reach the standard of care any reasonable individual would take; by failing to do so, it caused harm(s) to the victim.

⁷If a person has not committed any criminal acts yet (despite their mindset), it would be extremely onerous to suggest they should be punished or have their liberty restricted in any way due to thoughts that go through their mind on which they have not acted. Otherwise, this would be reminiscent of the *Thought Police* in George Orwell's *Nineteen Eighty-Four*, who punishes thoughtcrime.

Returning to the general predictive framework, given individual brain/behaviour data x and labelled outcome y , one can use AI/machine learning algorithms to develop (via training and test and cross-validation) a model $M : M(x) \rightarrow y$ that maps brain/behaviour data x onto outcome y . Moreover, one can use this trained model M to assess outcomes given data from a new subject x^{new} ; in other words, $y^{new} = M(x^{new})$. If the model captures the relationship between brain and behaviour data and outcomes in a general population, and the data are of good quality, then the estimated outcome y^{new} is expected to be close to its unobserved (but true) outcome. This may help assess potential outcomes during legal investigations when the outcomes are not directly/yet available.



(a) The triangle between the brain, behaviour, and law. One can either use brain data directly to assist legal judgement; one can also use behaviour data to assist legal judgment (because behaviours are outputs of the brain). **(b)** Brain decoding. Individual brain data are coupled (e.g., via point-wise multiplication) with trained brain-wise weights to yield outcomes of legal interests. **(c)** Behaviour decoding. Individual behaviour data are coupled with trained behaviour-wise weights to yield outcomes of legal interests. **(d)** Three types of outcomes of legal interests. Top right: Categorical determination (e.g., whether someone acted intentionally or unintentionally). Middle right: Continuous judgement (e.g., how severe is the crime). Bottom right: Future forecast (e.g., what is the likelihood of someone committing a crime in the future).

Fig 2: A schematic representation showing how predictive models developed on brain and behaviour data may assist legal decision-making.

The outcome of legal interest is not restricted to univariate cases (*i.e.*, with only a single outcome). By evaluating the brain patterns across subjects, machine learning algorithms can identify neural markers related to several outcomes of legal interest. When brain data are unavailable, behavioural data (such as speech patterns) recorded semi-continuously, for example, from wearable computers, can serve as surrogate behavioural markers.

On the other hand, if one could legally gather the data (*i.e.*, in line with the GDPR), it could be helpful to assist with knowing which persons of interest in the community are so they can be surveilled, like the 'heat list' in the US. MI6 has a similar list in the UK. See (Fazel et al., 2019) for a study on the prediction of violent reoffending in prisoners and individuals on probation.

Recently, using these technologies, one has begun to assess and predict abnormal mental states (Reinen et al., 2018), psychological episodes (Cao et al., 2018), and the presence of dementia (Teipel et al., 2008) in neuroscience and psychology. These explorations have proved the concept of and provided technical foundations for making categorical (present of illness or not) or continuous (how severe the symptoms are) enquiries to determine insanity, culpability, or ability to form intent. In principle, too, these technological advances and neurobiological findings can be translated to criminal law, especially in the terrain of outcome prediction.

But little do we know to what extent they can be applied and adopted in criminal law practice. A beginning can perhaps be made by testing these approaches using existing data that have been labelled (*i.e.*, the outcome such as intention to kill has been confirmed) and applying these approaches in several areas in criminal law to (a) find which types of data may yield accurate performance; (b) pinpoint which sub-fields of criminal law may at present benefit the most from such explorations; (c) identify areas where the models are not performing well and investigate the gaps. Another attempt may be to discuss and explore the possibilities of performing predictions using ensembled data (in other words, combining both brain/behaviour information and traditional biological data such as fingerprinting and DNA in criminal law) (see Section 3).

3. On detecting lies, bias, and visits to crime scenes

Central to robust and fair jurisprudence is accurately and consistently distinguishing punishable defendants from innocent ones. Importantly, for human judges, this requires reliable and reproducible methods to separate, in terms of closed-form (yes or no) answers, truthful statements from partially true or completely wrong statements given by the defendants.

In cross-examination, the counsel questioning the witness must ask ‘closed’ questions (*i.e.*, questions that will only elicit a yes or no response). For example, “you were in Howard House on the night of Friday, 26 November; please answer only yes or not?” Oftentimes, however, when faced with mounting legal documents and testimonies, it may be challenging to ascertain the trustfulness and unbiasedness of the evidence/statements. Machine learning may help to perform data reduction and select features to make dichotomous predictions.

One interfacing area between machine learning and brain science that may contribute to criminal law is crime scene detection. J.D. Haynes and colleagues at the Berstein Center in Berlin have been working on projects involving participants touring various virtual reality (VR) houses. After viewing these houses, the participants’ brains are scanned. Early results show that it is possible to identify the houses one had before been to based on his/her brain data. This shows the promise of using brain imaging techniques to determine whether a suspect had been to a crime scene (Smith, 2013).

Although such explorations (and the methodologies implemented) demonstrated the possibility of making dichotomous predictions using brain data, they are not without limitations. One difficulty is distinguishing whether the suspect had been to the crime scene while committing the crime or s/he was there accidentally. Another complication is determining the threshold of the prediction. Oftentimes, predictive algorithms for dichotomous outcomes yield a (predicted) score ranging from 0 (the suspect was predicted to be *surely not there*) and 1 (the suspect was predicted to be *surely there*). One could subsequently set a threshold, say, 0.6, such that an individual with a predicted outcome above the threshold is judged to be at the scene, and one with a predicted score below the threshold is judged not at the scene. It is, therefore, not only a prediction problem but also a legal (and perhaps a philosophical) one to determine a suitable threshold (as the threshold decides the sensitivity and specificity of the results).

Naturally, one would ask, at what stage(s) of the legal investigations should machine intelligence, brain, and behaviour data-based assessment and prediction be (more) involved? My views are as follows. During evidence collection, one may employ them when the witness is being interviewed by the lawyers preparing their witness statement (*i.e.*, weeks or months before the trial). Once in the courtroom, the witness will be reading the witness statement prepared by someone else, which itself may tamper with their brain signals (*i.e.*, if something is not worded exactly how they would have done), or they will only be answering yes/no questions. We cannot see at present a clear scenario where a witness (except perhaps the defendant) would agree to have one's personal data collected and processed, in the same vein that lawyers and jurors would not agree to this. More studies, discussions, and debates are needed regarding this (see [Section 5.6](#) for a further discussion).

Important to social interaction are trust and cooperation. They are integral elements in maintaining and facilitating effective and truthful communication between the defence lawyers and the accused, the plaintiff and the accusers, and judges and jurors. Some neuroscientists and neuroeconomists are trying to explain the neurological and behavioural underpinning of trust and cooperation. What has not yet been well explored, but is important to the law in general and criminal law in particular, is to assess and predict trust and cooperation using the brain data ([McCabe et al., 2001](#)). By directly examining the neural correlates of trust and cooperation, one may begin to find optimal strategies (and awards) needed to evaluate (and promote) trust (and cooperation) between different parties during legal investigations.

Brain signals are generally not subject to fabrication or distortion (but see [Section 5.2](#)). Social pressure may prevent one from expressing unpopular or socially unacceptable views ([Fazio et al., 1995](#), [Nosek and Banaji, 2002](#)). Brain data preserve objective measurements of one's thoughts and may provide alternative views regarding a suspect than potentially biases from the jurors' ⁸ ([Korn, Johnson and Chun, 2012](#)) (see also the privacy and ethics concerns in [Section 5.1](#)). Brain imaging data are also beginning to show promises to advance lie detection. For example, two companies, No Lie MRI and Cephos, are presently offering fMRI-based lie detection services in the criminal justice context ([Farah et al., 2014](#)) (see issues below). Taken together, coupled with brain and behaviour data, analytical tools may potentially help to identify and correct biases and false statements during evidence collection and court debates.

Despite promises, brain and behaviour data-based detections of a lie, bias, and visits to a crime scene at present face several major shortcomings. The **first** is regarding accuracy. Although impressive - given the mostly noninvasive nature and relatively low resolution of the data, the accuracy of image-based lie detection is not high ([Farah et al., 2014](#)). This may present a major problem for a practical adaptation of using brain and behaviour data in the context of criminal law at present. To be convicted, the legal standard is that the judge/jury must consider the defendant committed the crime "beyond a reasonable doubt". This is a very high standard of certainty that the defendant committed the crime; if any doubt exists, the defendant must be acquitted. In fact, for a murder trial, for instance, the defendant will face losing their liberty for many years. Therefore, for machine intelligence-based evidence to be useful, it must be highly accurate. The **second** is regarding reproducibility. Although laboratory and field tests have shown that detection results based on brainwave (*e.g.*, P300-MERMER) stimulated by words/pictures relevant to a crime scene are laudable ([Farwell, 2012](#), [Farwell and Smith, 2001](#)), it is possible to manipulate one's thoughts to conceal information during brain data-based detection ([Hsu et al., 2019](#)). In parallel to continuing to improve the identification accuracy by designing better predictive algorithms and developing

⁸Here, I assume that juror's brain data could also be analysed – likely jurors would not consent to this. Further work may discuss the possibility and usefulness of it in the future and, if useful, what is needed to make it possible.

improved data collection methods, one potential way to address these issues is to consider combining brain imaging data with other traditionally used physiological and behavioural data, such as one's facial expression (Ekman and Friesen, 1969), tone (Scherer et al., 1985), and content of speech (Vrij, 2008) via data fusion and ensemble-learning to assess, for example, lies, during legal proceedings (rather than using an isolated type of data).

4. On general brain and behaviour decoding

As science, technology, and law interface and integrate, it is perhaps worthwhile to sail into the treacherous water to discuss how the brain decoding of general thoughts, intentions, *mens rea*, and “automate” process ⁹, as well as behaviour-based prediction, may assist (or threaten) law and criminal justice (Vilares et al., 2017).

4.1. The need

There is a need to extend brain and behaviour decoding to aspects beyond behaviour prediction and lie detection. Here, I define **brain and behaviour decoding** as using brain and behaviour data to assess *general* outcomes of legal interest. Although lie and intention detection indeed fall into the general concept of brain decoding, their functions and goals are relatively specific (*e.g.*, to tell whether the defendant is lying). Legal decision-making, however, is sometimes indirect and often requires assessment and predictions beyond lie and intention detections. For example, a murder charge can sometimes be reduced to a charge of manslaughter if the defendant lost control (Laver, n.d.) ¹⁰. Here, brain decoding may contribute to such a case by showing, via brain and behaviour analysis, that the defendant's mental state has been *continuously* disrupted (see Section 5.3 for further discussion).

4.2. Technical promises

Experimental research in laboratories has demonstrated that it is possible to use brain data to unfold natural images and subjective contents we see (Naselaris et al., 2009, Kamitani and Tong, 2005), movies scenes we watch (Nishimoto et al., 2011), hidden intentions we have (Haynes et al., 2007, Haynes, 2011), and episodes of dreams we undergo (Horikawa et al., 2013). These lines of evidence and methods suggest the possibility of potentially reconstructing (crime) scenes one had witnessed and wrongful actions one had taken by showing images of the scenes and actions and examining their corresponding neural activities. Additionally, by establishing these findings, they have provided pipelines to test these possibilities.

Additionally, the ever-improving classification models, predictive algorithms, and longitudinal methods provide analytical foundations for assessing the three types of outcomes of legal interests in Section 2. The recent development of artificial neural networks may also provide novel insights into identifying complex hidden associations between mental data and outcomes of the legal interest (Arrieta et al., 2020); also see predictability vs. explainability in Section 5.5.

⁹One performs an action that typically requires cognition without much thought after performing it many times (Heiner, 1983).

¹⁰Note that *provocation* is no longer a defence to murder in the UK (Coroners and Justice Act (2009, c.25)) <https://www.legislation.gov.uk/ukpga/2009/25/section/56>; but the *loss of control* is still a defence to murder.

4.3. When free will is interplacated between brain decoding and law

Although the efficacy of brain- and behaviour-based prediction on lie and intention needs to be improved, perhaps many will agree with the existence of lie and intention, their neural correlates, and their implications in law. The existence of the free will, however, has not been universally agreed upon (Hume, 1748, Nichols, 2011, Heisenberg, 2009, Soon et al., 2013, Haynes, 2011, Bode et al., 2011, Soon et al., 2008, Haynes et al., 2007). In brief, free will states that one's brain (or mind) decides on its own rather than being governed by some deterministic rule. It nevertheless has a profound implication on the justice system.

My view on free will is an integrated one. On the one hand, I believe that free will exists in higher-order brain functions, including, for example, the PFC's involvement in moral reasoning (Greene et al., 2001, Heekeren et al., 2003, Goodenough, 2001, Moll, de Oliveira-Souza and Eslinger, 2003). On the other hand, I believe there also exists potentially deterministic operations in lower-order brain functions, such as colour and face recognition (Zeki and Chén, 2020). Other mental or cognitive faculties and their behavioural manifestations are perhaps generated by integrating the two; namely, they combine some deterministic neural processes and free will. For example, associating a white flag with surrender (combining deterministic colour recognition and the concept of surrender which is free will) and racial profiling (combining deterministic face recognition and prejudice which is arguably free will).

To promote discussion, suppose free will exists ¹¹. I proceed to discuss its impact on law and decision-making. Research has demonstrated the possibility of predicting free will via brain decoding (Wisniewski, Deutschländer and Haynes, 2019, Haynes, 2011). Yet it remains to distinguish criminal behaviour due to free will from criminal behaviour due to brain damage (which disrupts free will). The former would lead to more severe punishment. Relatedly, there is a possibility where free will is lurking in a criminal case, for example, when the PFC of the defendant is damaged. More specifically, some PFC damage may change behaviour significantly, resulting in violent and criminal behaviour (Brower and Price, 2001). But proving PFC damage in court does not necessarily exempt one from committing crimes. For example, minor PFC damage may lead to the inability to feel emotional pain. Between not feeling emotional pain and committing a crime potentially lurks the free will. More precisely, the minor brain damage leading to a lack of emotion (which is not necessarily crime-causing) and free will, which acts upon or in association with the lack of emotion, results in criminal behaviour. Predictive modelling is potentially helpful for assessing whether one 'lacks capacity'. If one commits a criminal act and evidence is provided (via brain decoding) that they lacked capacity at the time of the offence, then this can be a mitigating factor when sentencing them or considering how to rehabilitate them (Graham, 2019).

5. The challenges of applying brain and behaviour decoding in jurisprudence

5.1. Privacy, ethics, and prosecutorial abuse

Here, in light of recent development and debates, I discuss a set of concerns machine learning and AI-based legal enquires using neural and behavioural data may face regarding privacy, ethics, and prosecutorial abuses.

To begin, let's ask ourselves a few questions. Suppose you are taking a walk in a modern city where your faces have been scanned many times. Who possesses your face data? Who

¹¹ In the eyes of the law in the UK, it is considered that people are generally considered to have free will.

can analyse them? And are you clear to whom the results have been unveiled ¹² (Information Commissioner's Opinion, 2009, 2021)?

Now let's replace faces with brain images and behaviour measurements and cameras with imaging scanners and smartphones, and think again. Although remote, dynamic brain and behaviour decoding in public may require years' effort, brain and behaviour decoding, in general, has made notable progress in relatively stable environments, evidenced by works cited in this paper. With data acquisition and analytical methods advancing rapidly, it is, therefore, perhaps timely and necessary to discuss how to prevent a powerful (potentially unjust) "brain and behaviour decoder" from assessing one's brain signals and behaviour measurements recorded from hospitals, labs, and remotely at home, decoding one's thoughts and actions, and revealing them to others for undisclosed purposes.

The brain data can be considered as 'special category data' under the General Data Protection Regulation (GDPR) with regulations provided by the GDPR regarding how to process them and under what circumstances (Information Commissioner's Opinion, 2018). Nevertheless, "[c]ategorisation of brain-derived data is unclear in terms of the GDPR when it is not about health, nor stemming from medical devices" (Rainey et al., 2020). Perhaps more confusingly, if one day machines develop into becoming capable of "remote brain decoding" outside of laboratories, it is unclear whether they are permitted to record one's brain signals, whether they are allowed to decode the data (e.g., extraction and extrapolation), and with whom they can share the results.

As a part of personal data ^{13 14} that constitute of, and are largely ¹⁵ unique to an individual, brain data should, in my view, be protected in the same way identifiable health records ¹⁶ (such as disease status and severity) are protected. Indeed, recording brain data in hospitals and research laboratories at present requires strict ethical approval. Yet, there are some untravelled territories - where scholars working on AI/machine learning, brain and behavioural sciences, and law could potentially jointly lead the exploration and discussion. For example, can a prosecutor acquire (or demand to acquire) one's brain images freely and use them as court evidence (if so, under what circumstances; if not always, when not)? How can we prevent someone from abusing brain and behaviour data analysis to prosecute his/her opponent(s) and defend his/her client(s)?

¹²In recent years, the use of facial recognition data has been heavily regulated. The Information Commissioner's Office determines what personal data can and cannot be processed and gives out large fines if they consider that a person or entity has breached the GDPR. Companies such as Google, Marriot Hotels and British Airways have also received large fines from GDPR regulators, so there is a real incentive for companies to comply with the GDPR. For example, Amazon was recently fined £636 million for a breach of the GDPR: <https://www.bbc.co.uk/news/business-58024116>.

¹³The term 'personal data' refers to "any information relating to an identified or identifiable natural person ('data subject')"; an identifiable natural person "can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (see GDPR Article 4).

¹⁴See also GDPR Article 13: the subject needs to receive "meaningful information about the logic involved" in automated processing, and GDPR Article 22: the data controller needs to implement measures to ensure the data subject's rights, freedoms, and legitimate interests, for example, to contest the decision.

¹⁵There are shared functional and structural characteristics between brains, such as those in the primary and associated visual cortex.

¹⁶For example, "Data from BCIs and other brain recordings is often personal and may be as sensitive as health data" (Rainey et al., 2020).

Unlike personal information or digital records, which are relatively concrete ¹⁷, and encode information within a comparatively narrow and static scope ¹⁸, brain data are dynamic, variable, and contain limitless variability. By limitless, I mean there are perhaps a variety of ridiculous thoughts that may be going through our brain (which do not necessarily end up in verbal or behavioural outputs) from moment to moment. If one day we can relatively transparently decode the brain, these thoughts may become unnecessarily apparent to others in court. Imagine a “brain decoder” reveals that one thinks the jurors dress funny or are being unfair; would that not affect some of the jurors emotionally? Similarly, all stored memories, ridiculous or sensitive, are also at risk for others to see and potentially bias jurors’ views on the individual, in a way perhaps similar to jurors’ racial biases (Korn, Johnson and Chun, 2012).

At the extreme end, suppose there is one person who frequently thinks about committing (but does not practice) homicide. Decoding his/her thoughts may jeopardise his/her life and career and, due to stress and unwanted attention, may push him/her towards committing wrongdoings or crimes later in action rather than thoughts. Certainly, one could argue this person may have an underlying psychological or mental substrate and may be more likely (probabilistically speaking) to render crimes in the future, and therefore needs to be managed or treated; but this person currently has not committed anything unlawful and should not be, per law, punished.

Taken together, I argue that unless having the brain data available to the court will add additional, unbiased information towards making a better judgment and decision-making, one’s brain data should be treated as strictly as one’s private health data (such as data recorded from health apps on smartphones) following the GDPR, if not stricter (for reasons above). Additionally, further discussions are needed to clarify circumstances under which one’s brain (and behaviour) signals can be analysed to extract evidence without the individual’s, or, in cases where one cannot give one’s consent (*e.g.*, one has psychological and psychiatric illnesses), his/her family’s or physicians’ consent. Equally, consensuses need to be reached regarding situations where one’s brain (and behaviour) signals cannot be analysed, especially to clarify under what scenarios there needs strict prohibition or protection. A stringent guideline may protect cases where the brain and behavioural analyses may add little legal insights but cause one’s irregular brain activities to be unnecessarily revealed in the court (even just to his/her family and friends), potentially bringing stigma or prejudice to the individual’s daily life afterwards. Finally, although there have been clear regulations (Council of Europe, 1999) on when predictive tests can be made using genetic information, there have not been general regulations regarding the applications of predictive tests using brain and behaviour data, particularly in legal cases. There is, therefore, an urgent need to clarify the definitions, set the boundaries, and establish regulations (Chesterman, 2021a) regarding when and when not brain and behaviour data can be used in a predictive manner in legal cases.

5.2. Accuracy, reliability, and reproducibility

To date, brain- and behaviour-based predictions are largely based on models developed from laboratory settings where the conditions are controlled and noises reduced. Models and parameters developed under these circumstances may not reflect a broad range of sceneries

¹⁷One’s health and disease records are either written in physical forms or stored in digital forms; once recorded, they do not usually change significantly.

¹⁸Such as age and gender; they do not change as widely and wildly as individual (brain) thoughts.

in the real world or be generalisable to capture the wider variability in a population (especially concerning heavy-tail events ¹⁹). Such models, when applied to out-of-the-laboratory features and real-world legal cases, may become irreproducible. For example, in a lab, even when asked to consider countermeasures, a matched control may not strive to avoid a positive detection since no severer consequences are involved. But if a defendant is asked to make a statement (based on which one may receive a death sentence) and simultaneously has, knowingly, his/her brain signals recorded, s/he may make a vigorous mental effort to conceal the truth, such that the brain patterns are partially altered.

Recent research suggests that to robustly identify correlations with individual differences measures, fMRI studies will need thousands (if not tens of thousands) of participants (Marek et al., 2022). This implies that effect sizes concerning brain measures and variation in behaviour, personality *etc.* will be very small. When applied to legal cases, despite promises, the effect sizes - in machine learning-based predictive modelling in general and fMRI-based studies in particular - may be an order of magnitude more challenging than one anticipates.

5.3. Predicting future actions

Using brain imaging techniques, it is possible to show that someone's brain states are *periodically* disrupted (Reinen et al., 2018). But here I argue that demonstration of *mens rea* via brain decoding at present is perhaps restricted to a small group of individuals whose brain patterns are *continuously* disrupted/irregular. This is, in part, because of the difficulty of linking brain decoding to *actus reus* (criminal conduct).

More specifically, to convict someone of defence, it must be proved that the defendant had *mens rea* at the precise time of committing the *actus reus*. For example, suppose A wanted to kill B, and A was driving to B's house with a firearm with the intention of shooting and killing B when A got there. But on the way, a cyclist swerved in front of A's car (no fault of A's) and A tried but could not stop in time, and the cyclist was hit by the car and died. By chance, the cyclist was B. Even though A wanted to kill B and was driving over to his house intending to do so; he did not have the intention to hit the cyclist. In this case, A would not be guilty of murder because he did not have the intent to hit and kill the cyclist at the time A hit the cyclist. But A would have been guilty if A noticed B was cycling near A's car and A intentionally hit and killed B.

The point is that one has to have the intention to commit the precise act at the same time as s/he commits that act. It is unlikely (unless the person wears a wearable EEG daily) that we can scan the person's brain at the precise time of committing the act – typically only after the act. It is under investigation (*e.g.*, research done in the Haynes lab) the possibility to decode whether one had been to a scene (in essence, it is to decode past memory); it is, as far as I am aware, currently difficult to precisely decode (stored) memory. As such, even if brain decoding can show that one with psychosis has brain patterns that are periodically disrupted (Reinen et al., 2018); one cannot ascertain whether the intention of an act happened during the exact period. The data, however, may show a propensity to have a particular mental state; yet I highlight that this is not the same thing as proof of a particular mental state at a particular time. Nonetheless, a stronger argument can perhaps be made if brain decoding can reliably show that the subject's brain patterns are *continuously* disrupted, or whose disruption had surely covered the period during which the act was carried out.

¹⁹Events on a heavy-tail distribution very rarely happen (but do happen); as such, algorithms trained from samples may not capture those rare events. For example, suppose a neural lie detector was trained - using true and false statements as well as brain data - when giving these statements to healthy subjects and patients who have known-neurological disorders (such as psychosis) while having their brain data recorded. But suppose there is a subject with a unique type of brain lesion; this person does not belong to the healthy group or the psychosis group; as such, the algorithm may detect this person as lying given his/her brain data – even if s/he is not lying.

5.4. One model fits all vs. personalised prediction

I have, in most of this paper, assumed a population model for brain- and behaviour-decoding. In other words, we train a predictive model using both brain and/or behaviour data as inputs and labels (such as intention) as outputs from, say, 1,000 individuals, and we say that this model can be generalized to other, previously unseen subjects. A population model (if proven generalisable and reproducible) is helpful because it extracts, at the population level, the general relationship between brain/behaviour data and the outcomes of legal interest and, as such, can be extended to other, new individuals to make an assessment when labels (outcomes) are not available or observable.

Indeed, perhaps most would agree that there are shared population-level similarities between individuals (for both the functions and organisation of the brain and for types of crimes); it is nonetheless unreasonable to assume that a population-level model would suit every person. The reason is twofold. **First**, at the source of the data, not everyone's brain is the same. When developing brain (and behaviour) based prediction, at present, one typically standardises different brain images onto the same template (similarly, one can standardise the behaviour features). While such practice embraces the population elements of biology and has practical conveniences, it may miss important individual traits. **Second**, when mapping brain (or behaviour) data onto the outcomes of legal interest under a population model, it is assumed that the same pathways (or parameters) exist for every individual. Yet, it is not always true. For example, a population model discovers that, on average, the disruption of areas 1 and 2 leads to individuals having a specific outcome (say, intention to kill). Yet, there may exist alternative pathways (say areas 3 and 4, or areas 1, 2, 3, and 4) that also lead to the same outcome. There could be, in addition to the population map, some individual maps that deviate from the population norm but are equally important.

One way to address issues related to “one (population) model fits all” is to develop personalised legal predictive models that account for both population-level shared patterns and individual information. The idea is similar to personalised model development and personalised medicine in the healthcare sector ([Chén and Roberts, 2021](#)). As such, I do not expand on the technicality; rather, I highlight one potential difficulty: that is, it is difficult to obtain brain and behaviour data for each individual (especially in criminal law where individuals are perhaps not as willing to “donate” their biological data as patients do in health and medicine); equally difficult would be to develop a personalised model for each individual given the high-dimensionality of the parameters and complex combinations of criminal scenarios. A compromise can perhaps be made by developing sub-group models – developing a model for each specific sub-group of individuals (say, one model for homicide and another for burglary); one could additionally include *ad hoc* parameters capturing individual characteristics.

5.5. Predictability vs. explainability

Whereas it is important to continue to improve the accuracy of “black-box” models ²⁰ that yield accurate prediction (say that of intention), to convince the judges, juries and perhaps the public to adopt these models in legal investigations, it is also important to make them explainable.

Predictive models, however, are sometimes built at the cost of explainability. For example, regularisation methods reduce estimation variance but introduce bias, making the model less explainable. Ensemble methods improve overall predictability by averaging predictions from

²⁰Models that take in brain and behaviour data and produce labelled outcomes such as predicted intention, but the intermediate modelling steps are obfuscated or difficult to explain.

individual models; meanwhile, the ensembles become difficult to explain (Shmueli, 2010). Neural networks may uncover hidden associations between features and outcomes and yield accurate predictions, but most are as-of-yet difficult to explain.

One potential way moving forward is to combine expertise from machine learning, law, and biological science communities to work on explainable models in legal investigations. Recent years have seen great effort in making AI/machine learning models explainable (Arieta et al., 2020). There are also discussions of explainable machine intelligence (XAI) in specific fields such as medicine (Tjoa and Guan, 2020). Yet it remains to extend XAI to the field of criminal law by incorporating insights from lawyers, judges, and juries. This is, in part, because the data to be trained are domain (law) specific, and to make the parameters interpretable, one needs insights from experts in biology, law, and machine learning.

For example, suppose a group of machine learning experts find the parameters corresponding to one set of brain areas consistently associated with a certain type of intention; they need to, on the one hand, verify these findings with neurobiologists whether that area is indeed associated with intention and perhaps decision making (therefore the model makes biological sense), and on the other hand, consult the findings with law-experts whether such explanations makes legal sense and whether evidence derived as such can be used in court.

In return, biologists and law experts may inspire machine learning/AI scientists to develop more targeted models and methods driven by specific biological/legal problems. One example would be to have biologists and law-experts help to find and narrow down more concrete legal cases where the relationship between the input and output is relatively clear (for example, to find out cases where a relatively clear classification exists between the sentenced and not sentenced) and the traditional legal classification labour-intensive; one can then develop algorithms to estimate such a relationship in an automated way and use it as a baseline model to which future models could compare, and from which better (more accurate and/or more explainable) models can be developed. Note that training a (even simple) model on clear cases/data does not mean the application of the model is restricted or the approach is not novel. In my view, as long as the model can be deployed to make suitable predictions on new subjects, it is novel, and, if it addresses a prediction problem in a traditionally labour-intensive case, useful. Another example is to apply neurobiological and legal insights to reduce (redundant or less useful) data. For example, suppose, based on domain knowledge, one assumes areas outside of the PFC, hippocampus, and the visual cortex are not significantly involved in predicting whether a suspect has been to a crime scene. Machine learning/AI scientists can then develop models using data from only the PFC, hippocampus, and the visual cortex or penalise parameters associated with areas outside of these critical regions to make a better assessment instead of assigning weights across the entire brain ²¹.

5.6. When can brain data be used as court evidence for recommendations, sentences, and verdicts?

At present, there are several hypothetical occasions one may consider using brain- and behaviour-based findings in justice via machine learning.

The **first** is when the individual is likely a suspect or a potential re-offender, but there is otherwise no telling evidence. By inquiring into the brain and behaviour data, one may gain additional insights into the case. A recent Dutch study shows that it is possible to predict violent reoffending in prisoners and those on probation using predictive algorithms and routinely collected datasets used by criminal justice agencies (Fazel et al., 2019). Yet, as argued about, extreme caution needs to be made to guard against misjudgment.

²¹ Although the weights for areas outside of the PFC and visual cortex could be very small, it inevitably hinders explanation.

The **second** is when there is reasonable evidence suggesting a high likelihood that one may have (or have not) committed a crime, but no verdict has been achieved. With court approval and an individual's (or his/her physician's) consent, results from brain decoding can be used to help seek further evidence (if the likelihood for committing a crime is high) or as an additional piece of insight (if the likelihood for not committing a crime is high) (see "beyond a reasonable doubt" above).

A **third** is that once a verdict or sentence has been made, one examines, retrospectively with consent, brain- and behaviour-based evidence (which had not been used in legal proceedings and decision-making) to verify the decision and to test the performance of brain- and behaviour- based predictive algorithms (*e.g.*, had we made decisions using the brain and behaviour data, would we have come to the same conclusion?). This, together with data accumulation (algorithms are less accurate or generalisable if trained on limited data even if the data are of high quality), would help develop more accurate and reproducible methods to facilitate future scientific, technical, and legal investigations.

6. Final remarks

The creativity of the human brain has in the past helped to write, amend, and defend the law. Recent advances in machine learning have not only introduced the concept of brain and behaviour decoding but also provided examples for it. Today, standing on the shoulders of machine learning, brain and behavioural sciences, and law, the analyses of brain and behaviour data have begun to offer a glimpse into one's mental status, intention, and thoughts and how they may affect one's decisions and actions.

Nonetheless, brain decoding is still in its infancy; many exciting ideas need to be explored and tested. In this paper, I have discussed three important areas where advancements in machine intelligence and brain and behaviour studies may facilitate criminal law: to examine mental illness, evaluate insanity, and assess behaviour; to detect lies, biases, and visits to crime scenes; to decode one's thoughts and intentions.

Despite promises and advances, much work is needed, in the laboratories and, importantly, outside of the laboratories and in courts, to demonstrate and improve the efficacy and rigour of brain- and behaviour-based assessment, prediction, and decision-making in criminal law. Equally important is to formulate ethical, practical, safe, and reproducible procedures and protocols regarding recording, analysing, sharing, and extrapolating individual brain and behaviour information in legal practices. A beginning can perhaps be made by joining efforts from experts in machine learning, brain and behavioural sciences, and law. I hope that the observations, arguments, and hypotheses I have made, potentially controversial, may sprawl further discussions in evidence gathering, ethics, data possession, security, and privacy, and in testing the potential of employing brain and behaviour decoding in the future legal investigations.

Funding

Non-declared.

REFERENCES

- ANDERSON, A. K. and PHELPS, E. A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* **411** 305–309.
- ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCÍA, S., GIL-LÓPEZ, S., MOLINA, D., BENJAMINS, R. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58** 82–115.

- BODE, S., HE, A. H., SOON, C. S., TRAMPEL, R., TURNER, R. and HAYNES, J.-D. (2011). Tracking the unconscious generation of free decisions using ultra-high field fMRI. *PLOS One* **6** e21612.
- BROWER, M. C. and PRICE, B. (2001). Neuropsychiatry of frontal lobe dysfunction in violent and criminal behaviour: a critical review. *Journal of Neurology, Neurosurgery & Psychiatry* **71** 720–726.
- CAO, H., CHÉN, O. Y., CHUNG, Y., FORSYTH, J. K., MCEWEN, S. C., GEE, D. G., BEARDEN, C. E., ADDINGTON, J., GOODYEAR, B., CADENHEAD, K. S. et al. (2018). Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. *Nature Communications* **9** 1–9.
- CHANDRAN, R. (2022). As Malaysia tests AI court sentencing, some lawyers fear for justice. <https://www.reuters.com/article/malaysia-tech-lawmaking-idUSL8N2HD3V7>.
- CHÉN, O. Y. and ROBERTS, B. (2021). Personalized healthcare and public health in the digital age. *Frontiers in Digital Health* **3** 26.
- CHÉN, O. Y., CAO, H., REINEN, J. M., QIAN, T., GOU, J., PHAN, H., DE VOS, M. and CANNON, T. D. (2019). Resting-state brain information flow predicts cognitive flexibility in humans. *Scientific Reports* **9** 1–16.
- CHESTERMAN, S. (2021a). *We, the robots?: Regulating artificial intelligence and the limits of the law*. Cambridge University Press, Cambridge, UK.
- CHESTERMAN, S. (2021b). Through a glass, darkly: Artificial intelligence and the problem of opacity. *The American Journal of Comparative Law* **69** 271–294.
- EKMAN, P. and FRIESEN, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry* **32** 88–106.
- FARAH, M. J., HUTCHINSON, J. B., PHELPS, E. A. and WAGNER, A. D. (2014). Functional MRI-based lie detection: Scientific and societal challenges. *Nature Reviews Neuroscience* **15** 123–131.
- FARWELL, L. A. (2012). Brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cognitive Neurodynamics* **6** 115–154.
- FARWELL, L. A. and SMITH, S. S. (2001). Using brain MERMER testing to detect knowledge despite efforts to conceal. *Journal of Forensic Science* **46** 135–143.
- FAZEL, S., WOLF, A., VAZQUEZ-MONTES, M. D. and FANSHAW, T. R. (2019). Prediction of violent reoffending in prisoners and individuals on probation: A Dutch validation study (OxRec). *Scientific Reports* **9** 1–9.
- FAZIO, R. H., JACKSON, J. R., DUNTON, B. C. and WILLIAMS, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology* **69** 1013.
- GOODENOUGH, O. R. (2001). Mapping cortical areas associated with legal reasoning and moral intuition. *Jurimetrics* 429–442.
- GRAHAM, M. (2019). Navigating the criminal justice system for those who lack capacity. <https://www.stoneking.co.uk/literature/e-bulletins/navigating-criminal-justice-system-those-who-lack-capacity>.
- GREENE, J. D., SOMMERVILLE, R. B., NYSTROM, L. E., DARLEY, J. M. and COHEN, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* **293** 2105–2108.
- GUPTA, R., KOSCIK, T. R., BECHARA, A. and TRANEL, D. (2011). The amygdala and decision-making. *Neuropsychologia* **49** 760–766.
- HAYNES, J.-D. (2011). Decoding and predicting intentions. *Annals of the New York Academy of Sciences* **1224** 9–21.
- HAYNES, J.-D., SAKAI, K., REES, G., GILBERT, S., FRITH, C. and PASSINGHAM, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology* **17** 323–328.
- HEEKEREN, H. R., WARTENBURGER, I., SCHMIDT, H., SCHWINTOWSKI, H.-P. and VILLRINGER, A. (2003). An fMRI study of simple ethical decision-making. *NeuroReport* **14** 1215–1219.
- HEINER, R. A. (1983). The origin of predictable behavior. *The American Economic Review* **73** 560–595.
- HEISENBERG, M. (2009). Is free will an illusion? *Nature* **459** 164–165.
- HORIKAWA, T., TAMAKI, M., MIYAWAKI, Y. and KAMITANI, Y. (2013). Neural decoding of visual imagery during sleep. *Science* **340** 639–642.
- HSU, C.-W., BEGLIOMINI, C., DALL'ACQUA, T. and GANIS, G. (2019). The effect of mental countermeasures on neuroimaging-based concealed information tests. *Human Brain Mapping* **40** 2899–2916.
- HUME, D. (1748). *An enquiry concerning human understanding: A critical edition* **3**. A. Millar, London, UK.
- KAMITANI, Y. and TONG, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* **8** 679–685.
- KORN, H. A., JOHNSON, M. A. and CHUN, M. M. (2012). Neurolaw: Differential brain activity for black and white faces predicts damage awards in hypothetical employment discrimination cases. *Social Neuroscience* **7** 398–409.
- LAVER, N. (n.d.). Loss of Control, Provocation and the Criminal Law. <https://www.inbrief.co.uk/court-proceedings/provocation-and-criminal-law>.

- MAREK, S., TERVO-CLEMMENS, B., CALABRO, F. J., MONTEZ, D. F., KAY, B. P., HATOUM, A. S., DONOHUE, M. R., FORAN, W., MILLER, R. L., HENDRICKSON, T. J. et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature* **603** 654–660.
- MCCABE, K., HOUSER, D., RYAN, L., SMITH, V. and TROUARD, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* **98** 11832–11835.
- MOLL, J., DE OLIVEIRA-SOUZA, R. and ESLINGER, P. J. (2003). Morals and the human brain: a working model. *NeuroReport* **14** 299–305.
- NASELARIS, T., PRENGER, R. J., KAY, K. N., OLIVER, M. and GALLANT, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* **63** 902–915.
- NICHOLS, S. (2011). Is free will an illusion? <https://www.inbrief.co.uk/court-proceedings/provocation-and-criminal-law>.
- NISHIMOTO, S., VU, A. T., NASELARIS, T., BENJAMINI, Y., YU, B. and GALLANT, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* **21** 1641–1646.
- NOSEK, B. A. and BANAJI, M. R. (2002). (At least) two factors moderate the relationship between implicit and explicit attitudes. [10.31234/osf.io/fv6bw](https://doi.org/10.31234/osf.io/fv6bw).
- COUNCIL OF EUROPE (1999). Convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine: Convention on human rights and biomedicine (ETS No. 164). <https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=164>.
- INFORMATION COMMISSIONER’S OPINION (2009). The use of live facial recognition technology by law enforcement in public places. <https://ico.org.uk/media/about-the-ico/documents/2616184/live-frt-law-enforcement-opinion-20191031.pdf>.
- INFORMATION COMMISSIONER’S OPINION (2018). Special category data. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data>.
- INFORMATION COMMISSIONER’S OPINION (2021). The use of live facial recognition technology in public places. <https://ico.org.uk/media/2619985/ico-opinion-the-use-of-lfr-in-public-places-20210618.pdf>.
- PHELPS, E. A. (2002). *The cognitive neuroscience of emotion*. W. W. Norton & Company, New York, USA.
- RAINEY, S., MCGILLIVRAY, K., AKINTOYE, S., FOTHERGILL, T., BUBLITZ, C. and STAHL, B. (2020). Is the European Data Protection Regulation sufficient to deal with emerging data concerns relating to neurotechnology? *Journal of Law and the Biosciences* **7** Isaa051.
- REBER, J. and TRANEL, D. (2019). *Frontal lobe syndromes*. Elsevier B.V., Amsterdam, Netherlands.
- REINEN, J. M., CHÉN, O. Y., HUTCHISON, R. M., YEO, B., ANDERSON, K. M., SABUNCU, M. R., ÖNGÜR, D., ROFFMAN, J. L., SMOLLER, J. W., BAKER, J. T. et al. (2018). The human cortex possesses a reconfigurable dynamic network architecture that is disrupted in psychosis. *Nature Communications* **9** 1–15.
- SAPOLSKY, R. M. (2004). The frontal cortex and the criminal justice system. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359** 1787–1796.
- SCHERER, K. R., FELDSTEIN, S., BOND, R. N. and ROSENTHAL, R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research* **14** 409–425.
- SHMUELI, G. (2010). To explain or to predict? *Statistical Science* **25** 289–310.
- SMITH, K. (2013). Reading minds. *Nature* **502** 428.
- SOON, C. S., BRASS, M., HEINZE, H.-J. and HAYNES, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* **11** 543–545.
- SOON, C. S., HE, A. H., BODE, S. and HAYNES, J.-D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences* **110** 6217–6222.
- STUSS, D. T. and BENSON, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin* **95** 3.
- TEIPEL, S. J., MEINDL, T., GRINBERG, L., HEINSEN, H. and HAMPEL, H. (2008). Novel MRI techniques in the assessment of dementia. *European Journal of Nuclear Medicine and Molecular Imaging* **35** 58–69.
- TJOA, E. and GUAN, C. (2020). A survey on explainable artificial intelligence (xai): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **32** 4793–4813.
- VILARES, I., WESLEY, M. J., AHN, W.-Y., BONNIE, R. J., HOFFMAN, M., JONES, O. D., MORSE, S. J., YAFFE, G., LOHRENZ, T. and MONTAGUE, P. R. (2017). Predicting the knowledge–recklessness distinction in the human brain. *Proceedings of the National Academy of Sciences* **114** 3222–3227.
- VRIJ, A. (2008). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior* **35** 1323–1336.

- WISNIEWSKI, D., DEUTSCHLÄNDER, R. and HAYNES, J.-D. (2019). Free will beliefs are better predicted by dualism than determinism beliefs across different cultures. *PLOS One* **14** e0221617.
- ZEKI, S. and CHÉN, O. Y. (2020). The Bayesian-Laplacian brain. *European Journal of Neuroscience* **51** 1441–1462.