The computational psychiatry of antisocial behaviour and psychopathy

Pauli, Ruth; Lockwood, Patricia

DOI:
10.1016/j.neubiorev.2022.104995

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Pauli, R & Lockwood, P 2022, 'The computational psychiatry of antisocial behaviour and psychopathy', *Neuroscience & Biobehavioral Reviews*, vol. 145, 104995. https://doi.org/10.1016/j.neubiorev.2022.104995

Neuroscience and Biobehavioral Reviews 145 (2023) 104995

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

The computational psychiatry of antisocial behaviour and psychopathy

Ruth Pauli [a,*], Patricia L. Lockwood [a,b,c,**]

[a] Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK
[b] Institute for Mental Health, School of Psychology, University of Birmingham, Birmingham, UK
[c] Centre for Developmental Science, School of Psychology, University of Birmingham, Birmingham, UK

ARTICLE INFO

Keywords:
Antisocial
Psychopathy
Conduct disorder
Oppositional defiant disorder
Attention deficit hyperactivity disorder
Machine learning
Computational modelling
Computational psychiatry

ABSTRACT

Antisocial behaviours such as disobedience, lying, stealing, destruction of property, and aggression towards others are common to multiple disorders of childhood and adulthood, including conduct disorder, oppositional defiant disorder, psychopathy, and antisocial personality disorder. These disorders have a significant negative impact for individuals and for society, but whether they represent clinically different phenomena, or simply different approaches to diagnosing the same underlying psychopathology is highly debated. Computational psychiatry, with its dual focus on identifying different classes of disorder and health (data-driven) and latent cognitive and neurobiological mechanisms (theory-driven), is well placed to address these questions. The elucidation of mechanisms that might characterise latent processes across different disorders of antisocial behaviour can also provide important advances. In this review, we critically discuss the contribution of computational research to our understanding of various antisocial behaviour disorders, and highlight suggestions for how computational psychiatry can address important clinical and scientific questions about these disorders in the future.

1. Introduction

Antisocial behaviours violate societal norms and the rights of others, often with very serious social and economic consequences for victims and wider society (Romeo et al., 2006). Severe antisocial behaviour occurs in as many as 10% of children (Nock et al., 2006) and 2–3% of adults (Moran, 1999). When these behaviours occur without obvious explanation (e.g., due to substance abuse, psychosis, brain damage or learning disability), they are recognised as symptoms of mental disorder, but research into the causes has been beset by controversy (Crego and Widiger, 2015; Millon et al., 2002; Shipley and Arrigo, 2001; Arrigo and Shipley, 2001). There is general agreement that pathological antisocial behaviour is usually accompanied by some combination of affective disturbance and impulsive aggression, but the relative weight given to these factors varies greatly (Crego and Widiger, 2015). Some researchers and clinicians have emphasised the centrality of underlying affective traits, such as lack of empathy and guilt, with antisocial behaviour being merely a symptom (Cleckley, 1976). Others have argued that antisocial behaviour is the core dysfunction, at least for pragmatic diagnostic purposes (Crego and Widiger, 2015; American Psychiatric Association, 2013). Still others have viewed affective traits and antisocial behaviour as co-occurring but separable facets of disorder, or as markers for distinct subtypes of disorder (Crego and Widiger, 2015; Hare and Neumann, 2008).

Furthermore, age-appropriate criteria are critical when diagnosing children versus adults who exhibit antisocial behaviour. Unlike other mental health disorders such as depression and anxiety, which can be diagnosed with the same label across childhood to adolescence and adulthood, antisocial behaviour has different diagnostic labels based on the age of the individual. These different labels are necessary because social norms are different for children and adults, and because of the potential for stigma and psychological damage from diagnosing personality disorders in young children, whose personalities are still developing. The end result of these debates about underlying psychopathology, combined with age-specific diagnostic criteria, is a proliferation of disorders, such as psychopathy, antisocial personality disorder, conduct disorder with or without limited prosocial emotions, and oppositional defiant disorder, which all share some common 'antisocial' features but differ markedly in other respects (Fig. 1). Some of these disorders originally reflected attempts to capture the same phenomenon

* Correspondence to: Centre for Human Brain Health, University of Birmingham, UK.
** Correspondence to: University of Birmingham, UK.
E-mail addresses: r.pauli@bham.ac.uk (R. Pauli), p.l.lockwood@bham.ac.uk (P.L. Lockwood).

https://doi.org/10.1016/j.neubiorev.2022.104995
Received 23 July 2022; Received in revised form 21 November 2022; Accepted 7 December 2022
Available online 16 December 2022

within different frameworks (e.g., psychopathy and antisocial personality disorder), while others are more conceptually distinct (Crego and Widiger, 2015). Externalising disorders without an antisocial component, such as attention deficit/hyperactivity disorder (ADHD), also share some characteristics with these disorders and are often comorbid (Bayard et al., 2020; Faraone et al., 1991; Hinshaw et al., 1993) (Fig. 1). Among these disorders, the symptoms that overlap to varying degrees include lack of remorse and guilt, lack of empathy, callousness, impulsivity, irritability and anger, irresponsibility, and criminal or age-inappropriate defiant behaviours (Fig. 1). In childhood, conduct disorder, ADHD, and oppositional defiant disorder (ODD) are often grouped together as 'disruptive behaviour disorders' based on important overlap in terms of aetiology and presentation (Latimer et al., 2012). Such overlap hints that there may be transdiagnostic mechanisms that can be identified across these disorders (Fig. 1). Finally, lack of affect also overlaps with other clinical conditions such as apathy and anhedonia with and without depression (Husain and Roiser, 2018). Overall, antisocial behaviour is linked to a plethora of diagnoses and symptoms in children and adults, and while each diagnosis has distinctive characteristics, it is unlikely that all these diagnoses map onto fully distinct underlying psychopathologies. Uncovering the latent mechanisms that drive these different component processes, and the extent to which related disorders can be distinguished, will be essential for progress in the field and ultimately for clinical treatment.

Computational psychiatry has the potential to move beyond the current challenges of the diagnosis-based approach. Computational approaches take the position that mental disorders can be characterised at multiple interdependent levels, such as cognition, behaviour and neurobiology, using advanced statistical methods that capture latent constructs (Huys et al., 2016; Adams et al., 2016; Hauser et al., 2019). These different levels of explanation do not neatly delineate accepted diagnostic categories (Huys et al., 2016). For example, two disorders related to antisocial behaviour could share a common transdiagnostic dysfunction, such as impulsivity or lack of empathy. This dysfunction could be quantified by a mathematical model of behaviour, for example one that incorporates an impulsive bias to initiate actions regardless of expected outcomes (Guitart-Masip et al., 2012; Pauli et al., 2022), which is then implemented in a specific brain network or area (Lockwood et al., 2020). According to this understanding, unique constellations of dysfunctions can be considered as distinct mental disorders, although the dysfunctions themselves occur on spectrums that overlap both with other diagnoses and with mental health. Computational psychiatry shares part of its approach with other recent advances in classification of psychiatric disorders, such as Research Domain Criteria (RDoC) (Research Domain Criteria (RDoC), 2022), in that it considers cross-disorder dimensional 'domains' of mental functions that can be described at multiple levels of explanation. However, computational psychiatry has a more specific focus than RDoC on the nature of mechanistic links between different levels of explanation, from cognition and behaviour to neurobiology (Adams et al., 2016).

Within computational psychiatry, uncovering areas of commonality and distinction between disorders can be approached from two angles. First, a data-driven approach applies machine learning techniques to large datasets, to identify features that can classify disorders accurately and predict treatment outcomes. For example, a machine learning classifier might be used to judge whether a diagnosis really does identify a unique constellation of dysfunctions (hence high classification accuracy based on those dysfunctions), or to evaluate whether additional features, outside of the accepted diagnostic criteria, could be useful markers of disorder. This approach is agnostic to the causes or mechanisms involved, relying instead on consistent and reliable patterns emerging from sufficiently large datasets (Huys et al., 2016). Although machine learning encompasses a very wide set of techniques, including clustering and regression (Flach, 2012), we focus here on machine learning classifiers because these are more obviously different from traditional methods and are commonly used in the framework of computational psychiatry (Huys et al., 2016; Adams et al., 2016). Second, a theory-driven approach uses mathematical models to test explicit

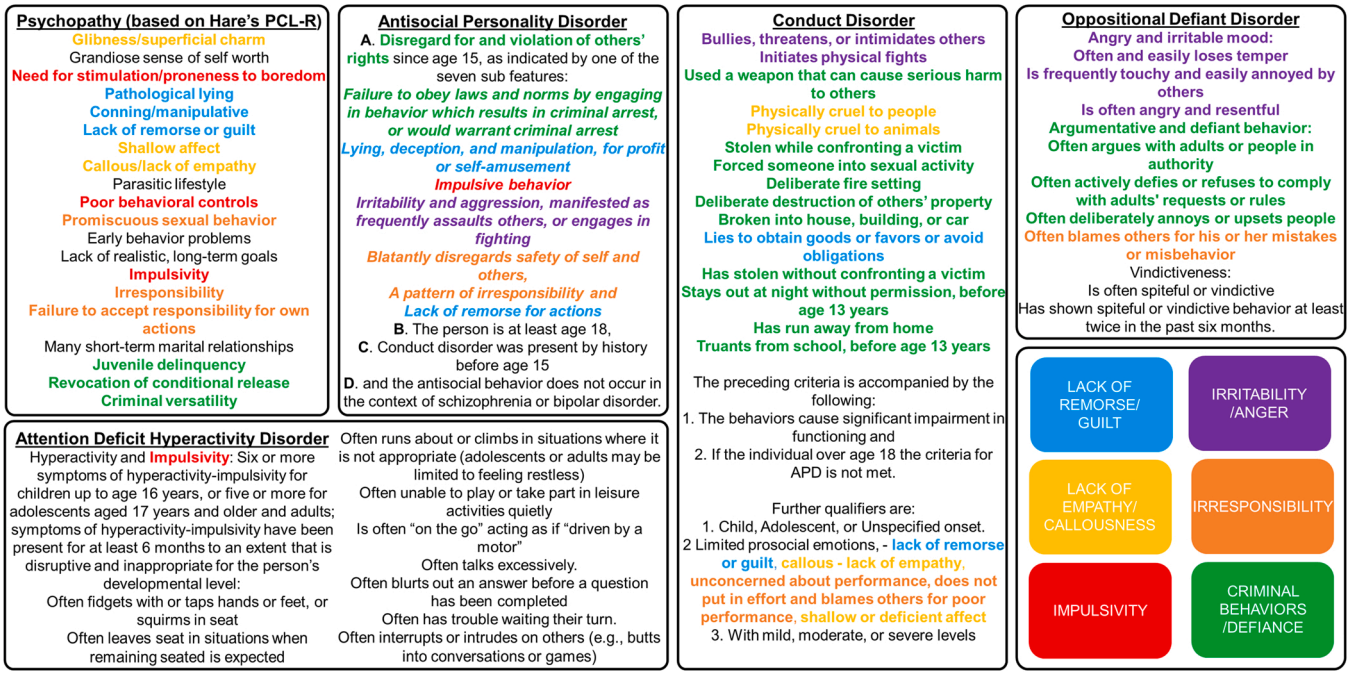

Fig. 1. Diagnostic criteria for disorders associated with antisocial behaviour and related disorders. Psychopathy, antisocial personality disorder, conduct disorder, oppositional defiant disorder, and ADHD share some overlapping features that could help to shed light on the component processes of each individual disorder. Transdiagnostic features such as lack of remorse/guilt, irritability/anger, lack of empathy/callousness, irresponsibility, impulsivity, and criminal behaviours/defiance could be particularly important for characterising computational and neurobiological mechanisms. The criteria listed here have been taken from the DSM-5 (with some rewording for brevity), except for psychopathy, which does not feature in DSM-5 and is instead captured by the Hare Psychopathy Checklist Revised (PCL-R).

hypotheses, commonly about cognitive-behavioural processes, but potentially incorporating multiple levels of explanation (Huys et al., 2016). Together, these two approaches can elucidate areas of overlap and difference between disorders, and identify latent cognitive mechanisms that drive behaviour at a transdiagnostic level. Although 'antisocial' diagnoses are an obvious candidate, computational work with these disorders is still at a stage of relative infancy (Brazil et al., 2018), and the antisocial construct may be further subdivided into more basic candidate mechanisms (Fig. 1). In this review, we aim to illustrate how computational psychiatry can further our understanding of antisocial behaviour and psychopathic traits across the lifespan. We first summarise the different diagnoses used to capture antisocial behaviour, highlighting several transdiagnostic clinical features that tend to occur across multiple disorders (Fig. 1). We then describe promising computational work in the field from both data-driven and theory-driven approaches, and conclude with suggested future directions for the field.

2. Diagnosing antisocial behaviour in children and adults

The psychiatry of antisocial behaviour has a long and complex history (Crego and Widiger, 2015; Millon et al., 2002; Shipley and Arrigo, 2001; Arrigo and Shipley, 2001; Bayard et al., 2020; Hervé, 2007). The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (5th edition; DSM-5 (American Psychiatric Association, 2013)) currently includes two diagnoses for severe and persistent antisocial behaviour. Conduct disorder is a behavioural disorder of childhood or adolescent onset, characterised by antisocial behaviour and violation of age-appropriate social norms (American Psychiatric Association, 2013). Criteria for diagnosis include physical and sexual violence, vandalism, and bullying (American Psychiatric Association, 2013). (Milder behavioural problems in children and adolescents, such as rule-breaking, defiance, and spiteful or vindictive behaviour, can instead be diagnosed as ODD (American Psychiatric Association, 2013)). Antisocial personality disorder is diagnosed in adults (age 18 +), and criteria include criminal antisocial behaviour with a history of conduct disorder, as well as impulsivity, manipulativeness, irresponsibility, and remorselessness (American Psychiatric Association, 2013) (Fig. 1). Adults who do not meet the full criteria for antisocial personality disorder can be diagnosed with conduct disorder if they meet the criteria, and if there is retrospective evidence for an onset before age 18 years. A similar diagnosis, dissocial personality disorder, is recognised by the World Health Organisation's International Classification of Diseases (World Health Organization, 2018).

Psychopathy is not included in the DSM-5, and it has sometimes, especially historically, been regarded as synonymous with antisocial personality disorder (Crego and Widiger, 2015; Buzina, 2012; Rogers and Rogstad, 2010). However, compared to antisocial personality disorder, psychopathy as a disorder is more closely associated with affective traits such as lack of conscience, callousness, and shallow affect (Cleckley, 1976; Hare and Neumann, 2008). It has therefore come to be seen as a separate and more severe personality disorder (Crego and Widiger, 2015; Millon et al., 2002; Buzina, 2012; Ogloff, 2006) (but see (Rogers and Rogstad, 2010)). As well as these affective traits, some researchers view fearlessness as an important feature of psychopathy, but its centrality to the disorder is still debated (Crego and Widiger, 2015; Lilienfeld et al., 2012; Lynam and Miller, 2012). In recent years there has been increasing recognition that psychopathic traits are not binary phenomena, and psychopathy can also refer to spectrums of affective and antisocial psychopathic traits associated with the personality disorder, but present to a lesser degree across the population (Crego and Widiger, 2015; Edens et al., 2006; Newman et al., 2005; Sellbom and Drislane, 2021) (but see (Coid and Yang, 2008)). Finally, the core callous-unemotional affective traits of psychopathy (low empathy, shallow affect, remorselessness, and lack of concern about performance in important activities) are now used to demarcate a more severe 'limited prosocial emotions' subtype of conduct disorder in the DSM-5 (American Psychiatric Association, 2013; Rowe et al., 2010; Kahn et al., 2012; Frick and White, 2008; Frick et al., 2003, 2005; Viding et al., 2005), although callous-unemotional traits are not binary phenomena and are typically somewhat elevated in conduct disorder regardless of subtype (Kliem et al., 2022) (Fig. 1). Similar phenomena are therefore described using different terminology in children and adults, a situation that is perhaps compounded by the tendency for most researchers to specialise in only one developmental period.

Overall, a number of transdiagnostic clinical features are present across related disorders. As well as antisocial behaviour, these transdiagnostic features include the callous-unemotional affective traits of psychopathy, lack of remorse or guilt, impulsive behaviour, aggression, irritability, and criminal behaviour or age-inappropriate defiance (Fig. 1). Some transdiagnostic features are essential diagnostic criteria for multiple disorders (antisocial behaviour), others are central to some disorders and more peripheral to others (e.g., callous-unemotional traits and lack of remorse or guilt), and some are commonly observed but of debated importance (e.g., fearlessness). Thus, there is a pressing need for the 'antisocial' diagnoses to benefit from the emerging framework and methods of computational psychiatry, as has already begun to happen for disorders such as anxiety and depression (Pulcu and Browning, 2017; Chen et al., 2015; Pike and Robinson, 2022; Brown et al., 2021). In this review, we focus on conduct disorder, antisocial personality disorder, and psychopathy, since computational work on ODD has, to our knowledge, not begun and because other excellent reviews of computational approaches to ADHD already exist (Latimer et al., 2012). However, because of the overlap between these different disorders, we highlight important studies where these are relevant to understanding common mechanisms.

3. Data-driven approaches

Machine learning classifiers use complex, multivariate data patterns to distinguish between different categorical classes (Flach, 2012). Classifiers are almost entirely data-driven in that, apart from selecting the data itself, the researcher has no further input into what variables the classifier should use. The advantage of this approach is that classifiers can detect combinations of features that together are highly predictive of a disorder or clinical outcome, but would escape detection by humans because of their complexity and the lack of a priori hypotheses relating to them. Conversely, precisely because of this data-driven flexibility, classifiers cannot distinguish between variables 'of interest' and variables that humans would regard as confounders in many contexts, such as age or ethnicity. Consequently, classifiers are most useful where there is a practical need to distinguish between groups, but where it is not essential to understand why groups differ.

Classifiers have been applied to antisocial behaviour disorders with some success. For example, Sato et al (Sato et al., 2011) distinguished between adults with antisocial personality disorder and psychopathy (Cooke et al., 1999) versus healthy controls with 80% accuracy, using voxel-level magnetic resonance imaging (MRI) grey matter volume data. Other researchers have used social media content (Asghar et al., 2021; Alotaibi et al., 2021; Wald et al., 2012; Henning, 2017; Sumner et al., 2012; Mahmud et al., 2021), Near Infra-red Spectroscopy (NIRS) (Dashtestani et al., 2019), speech patterns (Jain et al., 2019), videotaped head motion (Gullapalli et al., 2021), electroencephalogram (EEG) frequency data (Baumgartl et al., 2020), and histories of childhood abuse and caregiving (Schorr et al., 2021) to distinguish adults with psychopathy or antisocial personality disorder from healthy controls (see also Cope et al., 2014). Psychopathic traits have also been used as successful predictors of aggression (Suchting et al., 2018) and opiate and stimulant abuse (Ahn and Vassileva, 2016). Interestingly, one MRI-based classifier adopted a three-class approach based on severity of psychopathy, and performed better when moderate psychopathy was grouped with non-psychopathy rather than with severe psychopathy or as a separate class from both (Pearce, 2015). However, the criteria used

to define psychopathy were not explained. This is a common omission in papers where the focus is on machine learning methods rather than psychopathy, but given the complexity and debate surrounding psychopathy diagnoses, inclusion of this information would greatly increase the impact of computer science literature within computational psychiatry.

More generally, MRI-based classifiers could potentially suffer from low reliability, which was highlighted as a problem in the univariate literature by a recent meta-analysis of 134 MRI psychopathy studies (Deming et al., 2022). The reliability of multivariate MRI studies is currently unknown due to a lack of literature, but if similar problems do arise then these can be overcome with larger sample sizes and precise anatomical labelling (Deming et al., 2022). One important factor is that univariate and multivariate analysis can reveal unique findings and therefore any conclusions about regional atypicalities can consider both types of effect (Kriegeskorte et al., 2008). When working with functional data, it could also be the case that multivariate approaches are inherently more powerful as they can make use of multiple experimental conditions in parallel (Kriegeskorte et al., 2008).

In youths, a series of studies have been conducted with a sample of Chinese adolescent boys (aged 14–15) with 'pure' (non-comorbid) conduct disorder, and healthy controls. Different classifiers have achieved maximum accuracies ranging from 78% to 85% (Zhang et al., 2019, 2020a, 2018) using regional grey matter volumes from this sample, while resting state functional MRI data from a similar sample ('non-comorbid' Chinese adolescent boys aged 15–17) yielded accuracies of 75% (Lu et al., 2021) and 94% (Zhang et al., 2020b). These studies highlight the power of the multivariate machine learning approach. However, it should be noted that they were unusually homogenous in terms of age, sex, nationality, and lack of comorbid diagnoses, and therefore external validation on a separate dataset would be helpful.

In an unrelated, longitudinal sample of youths aged 9–10 years at baseline, a classification accuracy of 98% for predicting future conduct disorder was achieved using a combination of social, psychological, and biological predictors (Chan et al., 2022). These included age, sex, race, measures of neighbourhood safety, family income, parenting, baseline ADHD, oppositional defiant disorder, and conduct disorder diagnoses, tests of memory, attention, cognitive control, language, and reading ability, and global and local connectivity across multiple brain networks. Interestingly, the model was more accurate at identifying youths without conduct disorder (true negative rate) than youths with conduct disorder (true positive rate). This pattern has been observed with other classifiers (Pauli et al., 2021a, 2021b), and suggests that across a broad range of measures of psychological functioning, youths with conduct disorder are more likely to resemble healthy youths than vice versa. These findings highlight the importance of ensuring that diagnostic criteria strike an appropriate balance between sensitivity and specificity when diagnosing conduct disorder.

Perhaps more impressive, classifiers have also been able to distinguish between youths with closely related disorders. In a study of incarcerated teenagers, those with high psychopathic traits could be distinguished from those with low psychopathic traits with 69% accuracy using structural MRI data. This was much lower than when distinguishing either group from non-incarcerated controls, but still well above chance performance (Steele et al., 2017). Notably, however, these MRI-based classifiers were outperformed by an age-and-IQ-only classifier in the same sample. Likewise, youths with conduct disorder have been distinguished from youths with ADHD and youths with comorbid ADHD and conduct disorder. One study achieved 80% accuracy using electrocardiogram (ECG) data (Koh et al., 2022), and another achieved 98% accuracy using EEG data (Tor et al., 2021). Two studies have attempted to distinguish between conduct disorder with and without limited prosocial emotions. One study achieved 58% accuracy based on experiences of parenting (Pauli et al., 2021b), while a second achieved 52% (chance-level) accuracy based on facial emotion recognition abilities (Pauli et al., 2021a). Both of these studies reported superior performance (up to 75% accuracy) when distinguishing conduct disorder subtypes from healthy controls. Interestingly, in both papers, the true negative rate exceeded the true positive rate, indicating that youths with conduct disorder were more likely to be classified as controls than vice versa. As noted previously, this pattern has also been observed elsewhere (Chan et al., 2022).

Together, the data-driven literature concords with previous research pointing to neurobiological differences, albeit not universal, in people with antisocial behaviour disorders compared to healthy controls (Fairchild et al., 2013; Sterzer et al., 2007; Sebastian et al., 2016; De Brito et al., 2009, 2021a, 2021b). Although biology is an inextricable part of the causal chain for all human behaviour, including mental disorder, discussion of the biological correlates of antisocial behaviour can be especially controversial. This is because biological explanations for mental disorder are associated with reduced control and blame (Loughman and Haslam, 2018), and in the context of antisocial behaviour, such explanations can be seen either as excusing criminality (Morse, 1995) or unfairly stigmatising (Beltrán et al., 2021). These concerns are legitimate, but it is important that the degree of public sympathy or antipathy for people with mental health disorders is not allowed to override more objective interpretations of the literature. Thus, machine learning classifiers make an important contribution by demonstrating that biological differences can predict the presence of antisocial behaviour disorders in individuals with a high degree of accuracy. These findings, especially with further validation on external datasets, will hopefully reduce the sentiment that disorders that do not elicit sympathy are less 'real' than other disorders.

In summary, machine learning classifiers have demonstrated clear potential for identifying antisocial behaviour disorders across a wide range of data types, including some highly successful attempts to distinguish between closely related disorders using biological data (Koh et al., 2022; Tor et al., 2021). These successes raise the possibility of new biomarkers and risk markers being used to predict and detect disorders in the future. However, clinical use is still far from reality. First, an inherent limitation of the data-driven approach is that it is difficult, if not impossible, to understand the basis for successful classification. This is highlighted by the finding that in one sample, age and IQ were driving classification more strongly than regional grey matter volumes (Steele et al., 2017). Most clinicians would presumably find it unacceptable to diagnose based on age and IQ rather than disorder symptoms, but with classifiers it is always possible that demographic differences of this kind are actually the driving force behind classification success. This raises sensitive ethical issues around the clinical use of classifiers, especially regarding ethnic biases (Hitczenko et al., 2022). These ethical concerns have been addressed in detail elsewhere (Chen et al., 2021; Jurjako et al., 2019, 2020). Conversely, many researchers mitigate this issue by using demographically well-matched, numerically balanced groups, sometimes with variance associated with confounding factors already regressed out (e.g., (Sato et al., 2011); Zhang et al., 2020b; Pauli et al., 2021a; Steele et al., 2017). This increases interpretability and reduces opportunities for confounding variables to drive classification, but it is less data-driven and often results in samples that do not reflect the real-world incidence of health and disorder. Ultimately, classifiers must perform equally well on large, imbalanced, and messy real-world data before they can be of practical clinical interest beyond specialised research settings (Meeks, 2020).

Second, given the ongoing debate about the validity of the DSM-5 diagnostic approach (Casey et al., 2013), it is important to understand the extent to which closely related disorders are truly differentiable. Classifiers are ideally placed to address this debate because they can quantify how separable disorders are using a multivariate approach (Pauli et al., 2021a, 2021b; Steele et al., 2017; Koh et al., 2022; Tor et al., 2021). They have been successfully applied to other disorders, such as schizophrenia, with high accuracy (de Filippis et al., 2019), and we hope that multi-class, multi-disorder classification will become more

prevalent in the next few years. We also note that classifiers are only one form of machine learning, and other multivariate computational approaches, such as clustering, can also make valuable contributions in this area, although we have not focused on these techniques here (Coid and Yang, 2008; Cox et al., 2013; Vassileva et al., 2005; Swogger and Kosson, 2007; Hicks et al., 2004; Bronchain et al., 2020; Gong et al., 2022; Morana et al., 2006).

Third, it is important not to lose sight of the clinical relevance of research into antisocial behaviour disorders. At least for the foreseeable future, clinicians will generally treat psychological and behavioural symptoms directly, rather than intervening with biological correlates such as resting state connectivity. Consequently, data-driven approaches can only be the first step towards identifying relevant mechanisms for targeted clinical interventions. Rather than simply focusing on overall classification accuracy, researchers should use machine learning to identify combinations of features that are most predictive of specific disorders, or which can be reliably delineated across disorders. As previously described, different severities of psychopathic and callous-unemotional traits have been distinguished using classifiers (Pauli et al., 2021b; Steele et al., 2017). The other transdiagnostic features outlined here (Fig. 1) could be candidates for future successful machine learning approaches, especially regression-based methods that are more suited to handling continuous rather than categorical outcome measures.

4. Theory-driven approaches

In contrast to data-driven approaches, theory-driven approaches, such as computational modelling of behaviour, are strongly hypothesis-driven (Adams et al., 2016). To engage in theory-driven approaches, researchers must formulate precise mathematical models that are hypothesised to capture specific cognitive or biological processes, design experiments that engage these processes, and test how well the mathematical models capture the experimentally observed data (Adams et al., 2016; Lockwood and Klein-Flügge, 2020). A particular strength of this kind of modelling is its ability to capture latent or unobserved variables, such as learning rates or response biases, which can only be captured mathematically rather than through observing the data (Pauli et al., 2022; Lockwood and Klein-Flügge, 2020; Mars et al., 2012). Thus, while theory-driven modelling does not necessarily speak directly to the separability of different disorders, it can help to identify mechanisms that occur transdiagnostically across related disorders, as well as mechanisms that might distinguish between different disorders.

4.1. Reinforcement learning approaches to antisocial behaviour and psychopathy

Reinforcement learning theory describes how the learned reward or punishment value of an action influences the likelihood of repeating that action in the future (Sutton and Barto, 2018). Learning occurs when the outcome is unexpected, leading to a 'prediction error' that is then scaled by a learning rate and used to update the expected value of the relevant action for the future. Perhaps the clearest theoretical and empirical evidence for a transdiagnostic marker in antisocial behaviour and psychopathy is punishment learning. Deficits in learning from punishment have long been central to theoretical accounts of psychopathy, as well as related disorders such as conduct disorder (Newman and Kosson, 1986; Lykken, 1957; Blair et al., 2004a; Finger et al., 2011). However, it is only more recently that these learning processes have been modelled computationally. Typically, participants complete a learning task in which they must decide whether or not to respond to stimuli in order to gain rewards (positive feedback, points, or money) and avoid punishments. Responses are then modelled using a classic Rescorla-Wagner type model (Lockwood and Klein-Flügge, 2020; Rescorla and Wagner, 1972), in which possible actions are assigned expected values that are then converted into probabilities for engaging in that action (Lockwood and Klein-Flügge, 2020). The expected values are calculated from prediction errors, which are discrepancies between the actual and expected outcomes, and these in turn are scaled by learning rates, which determine how strongly the prediction errors affect the expected values (Lockwood and Klein-Flügge, 2020). In addition, these models contain a temperature or noise parameter that captures choice stochasticity. This is important for dissociating learning processes from behavioural variability in choosing the action with the highest expected value. A major strength of these models is that they are biologically plausible. This was demonstrated in the seminal work of Schultz and colleagues (Schultz et al., 1997), who observed that dopamine neurons in the ventral tegmental area carried a prediction error-like signal. The dopamine neurons increased their firing rate to an unexpected reward (a positive prediction error) and decreased their firing rate when expected rewards did not occur (a negative prediction error). Armed with a plausible biological basis and the ability to quantify latent learning processes, these models can reveal quantitative differences in learning speed that are associated with antisocial behaviour.

Using this approach, Oba and colleagues (Oba et al., 2019) investigated associations between punishment learning and self-reported psychopathic traits in Japanese undergraduates. The students completed a probabilistic go/no-go learning task, in which they had to decide whether or not to respond to stimuli to gain reward and avoid punishment. The different stimuli required either an active response to gain reward, an active response to avoid punishment, a lack of response to gain reward, or a lack of response to avoid punishment (Guitart-Masip et al., 2012). In the first of two studies using the probabilistic go/no-go task, participants were divided into high and low affective psychopathy groups. Compared to participants with low psychopathy scores, participants with high psychopathy scores exhibited poorer learning from the successful avoidance of punishment. Interestingly, when psychopathy was treated as a continuous rather than binary measure in the second study, this association was observed only for participants low in antisocial traits (Oba et al., 2019). This study suggests that impaired learning from punishment might be a cognitive computational process that is specifically impaired in those with high psychopathic traits. This fits with existing accounts of psychopathy, but had not been demonstrated computationally. Similar learning models have also been applied to learning about benefits for other people as well as for oneself. In a study of aging (Cutler et al., 2020), self-reported psychopathic traits were negatively associated with prosocial learning rates (i.e., learning rates for outcomes that help other people) in older adults, but not younger adults. These associations remained significant after controlling for group differences in the temperature parameter, suggesting specificity to learning rates. Together, these findings suggest that at least in healthy adults, psychopathic traits seem to be associated with learning rate differences.

Interestingly, reduced learning rates have also been observed in adults with ADHD, a disorder that shares impulsive features with psychopathy and antisocial personality disorder (Fig. 1). In a functional MRI (fMRI) study (Sethi et al., 2018), temporarily unmedicated adults with ADHD exhibited lower learning rates than healthy controls, were more likely to respond to novel stimuli, and exhibited heightened novelty signalling in the substantia nigra/ventral tegmental area. These group differences were much reduced when participants were given stimulant medication. However, ADHD is not always associated with reduced learning rates. Indeed, a review on computational approaches to ADHD suggested that the weight of evidence supports lower choice sensitivity in ADHD, but not lower learning rates (Ziegler et al., 2016).

Others have used different modelling approaches to investigate learning in psychopathy. In one study (Brazil et al., 2013), female participants had to decide which of two rectangles to select to gain rewards. In addition to learning the probability of reward for each rectangle, participants were given advice about which to choose on each trial. They could therefore be guided by the probability of reward and the probability of correct advice. These two factors were weighted using

computational parameters, which captured the relative influence of each type of information on choices. The use of reward probability in guiding decisions was positively associated with self-reported fearlessness and stress immunity, while the use of social information was negatively associated with self-reported stress immunity and positively associated with self-reported ability to manipulate others.

In another study using simulated data (Moul et al., 2021), callous-unemotional traits were reported to be associated with a longer 'learning window width'. This model differs from traditional reinforcement learning models (Lockwood and Klein-Flügge, 2020; Rescorla and Wagner, 1972) in that there is no calculation of a learning rate, or temperature parameter capturing choice noisiness/inconsistency. Instead, expected values are calculated by averaging across *n* previous trials, where *n* is the learning window width. A longer learning window is therefore roughly equivalent to a lower learning rate or slower forgetting (Moul et al., 2021), but whether a longer learning window directly reflects a lower learning rate, slower forgetting, or greater choice inconsistency is not distinguishable. By simulating data from a previously used learning task (Budhani et al., 2006), the authors demonstrated that callous-unemotional traits were associated with wider learning windows, and thus potentially (but not definitely) with lower learning rates. Taken together, these modelling studies suggest that psychopathic traits in adults are associated with poorer learning from punishment avoidance (Oba et al., 2019), poorer learning to help others (Cutler et al., 2020), and potentially slower learning in general (Moul et al., 2021), and that the specific traits of fearlessness and manipulativeness are associated with reward and social information use during learning (Brazil et al., 2013).

A smaller number of studies have modelled learning or learning-related processes in youths. One study investigated fMRI prediction error and expected value signals in youths aged 10–18 years with conduct disorder or oppositional defiant disorder (White et al., 2013). Participants viewed pictures of animals and could respond by pressing a button, or withhold responding. Some of the pictures were probabilistically associated with rewards and others with punishments when a response was made, while withholding responses resulted in neither reward nor punishment. Expected values for responding were calculated using prediction errors, with fixed learning rates across the whole sample. In youths with disruptive behavioural disorders compared to healthy controls, activity in the ventromedial prefrontal cortex was less strongly associated with expected values when deciding to respond, and activity in the anterior insula and caudate was less strongly associated with expected values when deciding to withhold responses. Additional associations were found for prediction errors during the outcome phase, with weaker associations between caudate activity and reward prediction errors relative to healthy controls, but stronger associations between caudate activity and punishment prediction errors (see also White et al., 2016).

A second study reported on fMRI measures of representational uncertainty in youths at high risk for antisocial personality disorder (Brazil et al., 2017). The youths were considered high-risk because they had all been arrested by the police for antisocial behaviour before age 12 years. The youths completed a behavioural task in which male faces were sometimes paired with electric shocks, but the punishment contingencies were unstable, so that the probability of getting a shock with a picture continually changed. The task was designed to capture contingency uncertainty (i.e., uncertainty about punishment contingency changes themselves) and change rate uncertainty (i.e., uncertainty about the rate of contingency changes during the task). Both forms of uncertainty were a challenge for successful learning, in which uncertainty must be minimised to learn about outcomes successfully (Brazil et al., 2017). In a factor analysis, activity in the left and right amygdala loaded onto the same factor as contingency uncertainty, and this factor was positively associated with callous-unemotional and impulsive-irresponsible psychopathic traits. In addition, activity in the left and right insula and the right amygdala loaded onto the same factor as change rate uncertainty, and this factor was positively associated with impulsive-irresponsible psychopathic traits. These findings suggest that both the affective and antisocial dimensions of psychopathy are related to neural representations of uncertainty during learning.

Taken together, these studies indicate that learning differences could underpin some of the latent constructs identified as contributing transdiagnostically to antisocial behaviour disorders. Reduced learning rates for punishing or rewarding stimuli could be consistent with an affective impairment that may be shared with other disorders. Although to our knowledge not investigated so far in antisocial behaviours disorders, models and tasks that capture a tendency to over-respond to stimuli regardless of outcome could characterise a transdiagnostic concept of impulsivity. Indeed, in typically developing children and adolescents, the tendency to respond to stimuli regardless of outcome (an action initiation bias) differs with age during the developmental period when conduct disorder often first develops (Pauli et al., 2022). Thus, learning models hold great promise for revealing latent component mechanisms (Adams et al., 2016).

4.2. *Value-based decision-making computational approaches to antisocial behaviour and psychopathy*

While fewer studies have focused on computational mechanisms beyond learning, research is beginning to uncover non-learning decision-making mechanisms as well. One such study (Lockwood et al., 2017a) tested people's willingness to exert physical effort (squeezing a hand grip) to obtain rewards for themselves and others. Participants tended to be apathetic about obtaining rewards for others, choosing to squeeze the grip less often than for themselves and exerting less force when they did so. This 'prosocial apathy' was elevated in people with the highest (subclinical) psychopathy scores. Indeed, prosocial apathy has also been linked to lack of empathy more broadly (Lockwood et al., 2017b). Another study compared sense of guilt and sense of fairness as explanations for non-reciprocal behaviour in an economic 'investor' game (Driessen et al., 2021) (see also Vieira et al., 2014). Fairness and guilt were modelled as weightings against decisions that were unfair (giving the investor an unequal share of the payoff) and guilt-inducing (giving the investor less than they would have expected from their investment). The fairness parameter and self-reported psychopathic traits were both associated with non-reciprocal behaviour (failure to pay dividends to an investor), although neither the fairness nor the guilt parameters were directly associated with psychopathy scores. However, in a similar study in adolescents (Yu et al., 2022), higher callous-unemotional traits were associated with lower aversion to getting more than one's fair share, but not to getting less than one's fair share. Similar patterns were observed in young adults in the same study (Yu et al., 2022), and another recent study with adults reported that psychopathy scores were associated with less anticipated guilt in a different economic investor (trust) game (Gong et al., 2019). Finally, one study investigated harmful and helpful behaviour to self and others (Contreras-Huerta et al., 2020). In this study, people with higher affective reactivity scores (calculated from self-report psychopathy measures) were more averse to harming others for profit, and more willing to exert effort to help others. Together, these studies suggest that, similar to learning rate differences, basic aspects of value-based decision-making might be altered in those with subclinical psychopathic traits. In particular, prosocial effort discounting and computations of fairness and harm could be important latent computational markers to focus on in future research.

In summary, the theory-driven computational literature on antisocial behaviour disorders currently resembles a 'patchwork', with numerous studies addressing different hypotheses, using different modelling approaches, in different populations. The small number of studies and the diversity of their approaches makes interpretation of the literature a challenge. Despite this, some broad patterns can clearly be discerned. First, computational modelling of reinforcement learning indicates that

psychopathy (Moul et al., 2021) and conduct disorders (White et al., 2013, 2016; Brazil et al., 2017) are associated with differences in latent learning processes, and these differences might be specifically related to punishment learning (Oba et al., 2019). Psychopathy is also associated with specific difficulties in learning contingency reversals (Moul et al., 2021; Budhani et al., 2006), which might reflect an impulsive difficulty inhibiting responses in a potentially rewarding environment (O'Brien and Frick, 1996) as well as punishment learning difficulties (Blair et al., 2004b). Previous computational work in healthy adolescents has suggested that impulsive responding can mimic reward sensitivity in the go/no-go learning contexts that are often used in psychopathy research (Pauli et al., 2022). Second, computational studies point to differences in how people with elevated psychopathy scores learn (Cutler et al., 2020) and make decisions that impact other people (Lockwood et al., 2017a; Contreras-Huerta et al., 2020), and suggest that insensitivity to fairness as well as apathy might be driving some of these differences (Lockwood et al., 2017a; Driessen et al., 2021; Yu et al., 2022). Third, computational modelling has been combined with neuroimaging to elucidate brain mechanisms connected with these learning and moral behavioural differences (White et al., 2013, 2016; Brazil et al., 2017), although this work is still in its infancy. Overall, however, the sparsity and diversity of the literature mean that we cannot as yet identify transdiagnostic markers of antisocial behaviour disorders with confidence. We suggest that a focus on identifying latent cognitive mechanisms linked to punishment learning and social processes could be fruitful from a transdiagnostic perspective and perhaps contribute more widely to understanding whether the various disorders associated with antisocial behaviour should be interpreted in terms of symptom severity or as distinct disorders. More broadly, from the common symptoms across different antisocial behaviour disorders, there are clear candidates for further investigation (Fig. 1).

Besides the general sparsity of the literature, there a few specific gaps that prevent a transdiagnostic understanding of antisocial behaviour. First, researchers have only very rarely tested people with antisocial behaviour disorders or equivalently serious behaviour, and then usually only in conduct disorder and oppositional defiant disorder (White et al., 2013, 2016; Brazil et al., 2017). With these few exceptions, there is a virtually exclusive focus on low-level, subclinical psychopathic traits in healthy volunteers (Oba et al., 2019; Cutler et al., 2020; Driessen et al., 2021; Yu et al., 2022; Contreras-Huerta et al., 2020). This contrasts sharply with the data-driven literature (Sato et al., 2011; Pearce, 2015; Zhang et al., 2019, 2020a, 2018, 2020b; Pauli et al., 2021a, 2021b) and conventional psychopathy research (e.g., (Blair et al., 2006); Gregory et al., 2015; Baskin-Sommers et al., 2010), where the focus is largely on diagnosed individuals. While there are good arguments for treating psychopathy as a set of traits rather than a category of disorder (Crego and Widiger, 2015; Edens et al., 2006; Newman et al., 2005; Sellbom and Drislane, 2021), a focus on purely subclinical phenomena may not generate transdiagnostic markers across different clinical disorders. Furthermore, these studies do not sufficiently address the more conventional understanding of psychopathy as a severely disabling phenomenon, whether or not it is regarded as a distinct disorder. Modelling focused on diagnosed individuals is therefore needed, to allow a clearer analysis of how computational work can elucidate cognitive mechanisms in severe antisocial behaviour. In parallel, diagnostic criteria could be revised to account for progress in capturing the basic mechanisms that drive these disorders and explain pathology across psychological and neurobiological levels.

Second, much of the computational work that touches on psychopathy has not had psychopathy as its primary focus (Cutler et al., 2020; Contreras-Huerta et al., 2020), and there are only a few examples of computational work by researchers studying psychopathy (e.g., Brazil et al., 2017). Perhaps as a result, differences in punishment and reward learning, which have repeatedly been linked to antisocial behaviour disorders (Blair et al., 2004b; Byrd et al., 2014), have rarely been modelled (Oba et al., 2019). This is despite such studies being quite common in other populations (Guitart-Masip et al., 2012; Pauli et al., 2022; Raab and Hartley, 2020; Palminteri et al., 2016, 2015). Thus, computational work by researchers with expertise in antisocial behaviour disorders is very much needed; in particular, more extensive work is needed on latent reward and punishment learning mechanisms in disorders such as psychopathy and conduct disorder. Recent work suggests that reward and punishment learning exhibit normative developmental differences from childhood to adolescence, with reward learning remaining stable while punishment learning rates increase (Pauli et al., 2022). This difference could be developmentally relevant, given that conduct disorder often emerges during adolescence, and could represent an important target for intervention.

Third, and related, it will be important to demonstrate how these latent mechanisms are linked to real-world behaviour. For example, it would be helpful to understand whether punishment learning rates can predict relevant real-world behaviour such as recidivism. Indeed, this may be an area of fruitful overlap between theory-driven and data-driven approaches, with latent mechanisms derived from computational models being used as predictors of disorder or outcome in machine learning classifiers.

5. Summary and conclusions

For many years, our understanding of antisocial behaviour disorders has been beset by scientific, ethical, and terminological controversies, resulting in a plethora of closely related disorders that are recognised within different psychiatric traditions, in different age groups, and with overlapping but non-identical diagnostic criteria (Crego and Widiger, 2015; Millon et al., 2002; Shipley and Arrigo, 2001; Arrigo and Shipley, 2001; Buzina, 2012). Debate continues as to how truly separable these disorders are (Crego and Widiger, 2015; Millon et al., 2002; Shipley and Arrigo, 2001; Arrigo and Shipley, 2001; Buzina, 2012), and computational psychiatry, with its focus on transdiagnostic mechanisms and classification of disorders, is well placed to address many of these controversies (Huys et al., 2016; Adams et al., 2016).

Data-driven approaches, with their 'agnostic' approach to distinguishing between different classes of disorder and health, have demonstrated that psychopathy, antisocial personality disorder, and conduct disorder can be reliably distinguished from health using a range of neuroimaging, behavioural, and questionnaire-derived measures (Sato et al., 2011; Zhang et al., 2019, 2020a, 2018, 2020b; Chan et al., 2022; Pauli et al., 2021a, 2021b). The successful identification of these disorders, based on criteria other than the diagnostic criteria (e.g., neuroimaging data), suggests that they all capture genuine psychopathology to a large extent. However, attempts to distinguish between closely related disorders or subtypes of disorders have been much rarer, and often less successful (Pauli et al., 2021a, 2021b; Steele et al., 2017). Whether or not closely related disorders, such as psychopathy and antisocial personality disorder, can be reliably distinguished from each other is an important empirical question, and one where data-driven computational psychiatry has much potential. If classifiers are able to distinguish between these disorders using their (hypothesised) key dysfunctions, then there is a strong argument for continuing to treat these disorders as genuinely clinically distinct phenomena. If this is ultimately not achievable, then these disorders should either be reconceptualised as capturing the same psychopathology, or better diagnostic criteria must be identified. Classifiers have also been used to distinguish between different severity levels within a diagnostic category (Pearce, 2015). This is another area which could contribute fruitfully to the clinical literature, by identifying 'cut-off' points where a set of dysfunctions become a disorder rather than 'normal variation' in the healthy population (Coid and Yang, 2008; Kimonis et al., 2014). Such an approach could also be adapted to investigate whether disorders such as ODD and conduct disorder, or antisocial personality disorder and psychopathy, are better conceptualised as distinct categories or variations in severity of the same underlying psychopathology. Although early

attempts at classification have naturally had to focus on methods development and proof-of-concept to some extent, addressing clinically relevant questions in an ethically appropriate way will become more important in the future if data-driven computational psychiatry is to have a real impact in mental health research.

Theory-driven approaches have generally focused on learning processes, due to the well-established mathematical models in this area (Lockwood and Klein-Flügge, 2020; Rescorla and Wagner, 1972). Although the literature is still quite sparse, psychopathic traits have been associated with differences in learning about outcomes for oneself (Brazil et al., 2013, 2017; Moul et al., 2021) and other people (Cutler et al., 2020), specific difficulties in learning from successful avoidance of punishment (Oba et al., 2019), and differences in moral cognition (Driessen et al., 2021; Yu et al., 2022), apathy (Lockwood et al., 2017a; Contreras-Huerta et al., 2020), and prosociality (Contreras-Huerta et al., 2020). Differences in learning mechanisms have also been observed in conduct disorder (White et al., 2013, 2016). Because the theory-driven literature is in its infancy, it is not yet possible to identify many latent cognitive mechanisms underlying transdiagnostic features of disorders with confidence. However, it seems likely that punishment learning deficits will be one such marker (Oba et al., 2019), and prosocial apathy is another candidate based on empirical evidence (Lockwood et al., 2017a, 2017b; Contreras-Huerta et al., 2020). Going forward, modelling studies with a central focus on psychopathy and antisociality will be important, as will research with more severely (and clinically) impaired participants, and research that specifically addresses the transdiagnostic nature of latent cognitive mechanisms in antisocial behaviour. More broadly, computational approaches hold promise for informing theoretical accounts of antisocial behaviour and psychopathy, in addition to uncovering core mechanisms (Moul et al., 2021; Prosser et al., 2018).

Finally, as the literature grows, there will be new potential to combine data-driven and theory-driven computational approaches to address questions that have not been answered by conventional research methods. For example, latent cognitive mechanisms underlying reward and punishment learning, moral reasoning, and prosocial apathy might be common to all antisocial behaviour disorders (and other disorders too), or occur in only a subset of these disorders. These mechanisms could then be used with machine learning classifiers to learn more about when disorders should be seen as separate phenomena, when they should be collapsed into broader categories, and when they might be better conceived of as a constellation of traits rather than as disorders in the conventional sense. Relatedly, those common features already established from current diagnostic criteria (Fig. 1) are also important candidates to investigate empirically and theoretically across antisocial disorders.

In conclusion, computational psychiatry has demonstrated the potential to address important clinical and scientific questions about antisocial behaviour and psychopathy. If this potential can be applied to the most pressing questions in the field, then the integration of computational expertise into antisocial behaviour research could pave the way for genuinely innovative and exciting new research into these disorders and the frequently transdiagnostic symptoms that comprise them.

Data availability

No data was used for the research described in the article.

Acknowledgements

R.P was supported by an ESRC post-doctoral fellowship award (ES/V011324/1). P.L. L was supported by a Medical Research Council Fellowship (MR/P014097/1 and MR/P014097/2), a Jacobs Foundation Research Fellowship, and a Sir Henry Dale Fellowship funded by the Wellcome Trust and the Royal Society (223264/Z/21/Z).

References

Adams, R.A., Huys, Q.J.M., Roiser, J.P., 2016. Computational psychiatry: towards a mathematically informed understanding of mental illness. J. Neurol. Neurosurg. Psychiatry 87 (1), 53–63. https://doi.org/10.1136/jnnp-2015-310737.

Ahn, W.Y., Vassileva, J., 2016. Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. Drug Alcohol Depend. 161, 247–257. https://doi.org/10.1016/j.drugalcdep.2016.02.008.

Alotaibi, F.M., Asghar, M.Z., Ahmad, S., 2021. A hybrid CNN-LSTM model for psychopathic class detection from tweeter users. Cogn. Comput. 13 (3), 709–723. https://doi.org/10.1007/s12559-021-09836-7.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Association.

Arrigo, B.A., Shipley, S., 2001. The confusion over psychopathy (I): historical considerations. Int. J. Offender Ther. Comp. Criminol. 45 (3), 325–344. https://doi.org/10.1177/0306624X01453005.

Asghar, J., Akbar, S., Asghar, M.Z., Ahmad, B., Al-Rakhami, M.S., Gumaei, A., 2021. Detection and classification of psychopathic personality trait from social media text using deep learning model. Comput. Math. Methods Med. 2021, e5512241 https://doi.org/10.1155/2021/5512241.

Baskin-Sommers, A.R., Wallace, J.F., MacCoon, D.G., Curtin, J.J., Newman, J.P., 2010. Clarifying the factors that undermine behavioral inhibition system functioning in psychopathy. Pers. Disord. Theory Res. Treat. 1 (4), 203–217. https://doi.org/10.1037/a0018950.

Baumgartl, H., Dikici, F., Sauter, D., Buettner, R., 2020. Detecting Antisocial Personality Disorder Using a Novel Machine Learning Algorithm Based on Electroencephalographic Data, p. 14.

Bayard, F., Nymberg Thunell, C., Abé, C., et al., 2020. Distinct brain structure and behavior related to ADHD and conduct disorder traits. Mol. Psychiatry 25 (11), 3020–3033. https://doi.org/10.1038/s41380-018-0202-6.

Beltrán, S., Sit, L., Ginsburg, K.R., 2021. A call to revise the diagnosis of oppositional defiant disorder—diagnoses are for helping, not harming. JAMA Psychiatry 78 (11), 1181–1182. https://doi.org/10.1001/jamapsychiatry.2021.2127.

Blair, K.S., Morton, J., Leonard, A., Blair, R.J.R., 2006. Impaired decision-making on the basis of both reward and punishment information in individuals with psychopathy. Pers. Individ. Differ. 41 (1), 155–165. https://doi.org/10.1016/j.paid.2005.11.031.

Blair, R.J.R., Mitchell, D.G.V., Leonard, A., Budhani, S., Peschardt, K.S., Newman, C., 2004a. Passive avoidance learning in individuals with psychopathy: modulation by reward but not by punishment. Pers. Individ. Differ. 37 (6), 1179–1192. https://doi.org/10.1016/j.paid.2003.12.001.

Blair, R.J.R., Mitchell, D.G.V., Leonard, A., Budhani, S., Peschardt, K.S., Newman, C., 2004b. Passive avoidance learning in individuals with psychopathy: modulation by reward but not by punishment. Pers. Individ. Differ. 37 (6), 1179–1192. https://doi.org/10.1016/j.paid.2003.12.001.

Brazil, I., Hunt, L., Bulten, B., Kessels, R., De Bruijn, E., Mars, R., 2013. Psychopathy-related traits and the use of reward and social information: a computational approach. Accessed April 27, 2022. https://www.frontiersin.org/article/ Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00952.

Brazil, I.A., Mathys, C.D., Popma, A., Hoppenbrouwers, S.S., Cohn, M.D., 2017. Representational uncertainty in the brain during threat conditioning and the link with psychopathic traits. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 2 (8), 689–695. https://doi.org/10.1016/j.bpsc.2017.04.005.

Brazil, I.A., van Dongen, J.D.M., Maes, J.H.R., Mars, R.B., Baskin-Sommers, A.R., 2018. Classification and treatment of antisocial individuals: from behavior to biocognition. Neurosci. Biobehav. Rev. 91, 259–277. https://doi.org/10.1016/j.neubiorev.2016.10.010.

Bronchain, J., Raynal, P., Chabrol, H., 2020. Heterogeneity of adaptive features among psychopathy variants. Pers. Disord. Theory Res. Treat. 11 (1), 63–68. https://doi.org/10.1037/per0000366.

Brown, V.M., Zhu, L., Solway, A., et al., 2021. Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy. JAMA Psychiatry 78 (10), 1113–1122. https://doi.org/10.1001/jamapsychiatry.2021.1844.

Budhani, S., Richell, R.A., Blair, R.J.R., 2006. Impaired reversal but intact acquisition: Probabilistic response reversal deficits in adult individuals with psychopathy. J. Abnorm. Psychol. 115 (3), 552–558. https://doi.org/10.1037/0021-843X.115.3.552.

Buzina, N., 2012. Psychopathy – historical controversies and new diagnostic approach. Psychiatr. Danub, 24(2), p. 9.

Byrd, A.L., Loeber, R., Pardini, D.A., 2014. Antisocial behavior, psychopathic features and abnormalities in reward and punishment processing in youth. Clin. Child Fam. Psychol. Rev. 17 (2), 125–156. https://doi.org/10.1007/s10567-013-0159-6.

Casey, B.J., Craddock, N., Cuthbert, B.N., Hyman, S.E., Lee, F.S., Ressler, K.J., 2013. DSM-5 and RDoC: progress in psychiatry research? Nat. Rev. Neurosci. 14 (11), 810–814. https://doi.org/10.1038/nrn3621.

Chan, L., Simmons, C., Tillem, S., Conley, M., Brazil, I.A., Baskin-Sommers, A., 2022. Classifying conduct disorder using a biopsychosocial model and machine learning method. Biol. Psychiatry Cogn. Neurosci. Neuroimaging. https://doi.org/10.1016/j.bpsc.2022.02.004.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., Kusumi, I., 2015. Reinforcement learning in depression: a review of computational research. Neurosci. Biobehav. Rev. 55, 247–267. https://doi.org/10.1016/j.neubiorev.2015.05.005.

Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M., 2021. Ethical machine learning in healthcare. Annu. Rev. Biomed. Data Sci. 4 (1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Cleckley, H., 1976. The Mask of Sanity, fifth ed. Mosby.

Coid, J., Yang, M., 2008. The distribution of psychopathy among a household population: categorical or dimensional? Soc. Psychiatry Psychiatr. Epidemiol. 43 (10), 773. https://doi.org/10.1007/s00127-008-0363-8.
Contreras-Huerta, L.S., Lockwood, P.L., Bird, G., Apps, M.A.J., Crockett, M.J., 2020. Prosocial behavior is associated with transdiagnostic markers of affective sensitivity in multiple domains. Emotion. https://doi.org/10.1037/emo0000813.
Cooke, D.J., Michie, C., Hart, S.D., Hare, R.D., 1999. Evaluating the screening version of the hare psychopathy checklist—revised (PCL:SV): an item response theory analysis. Psychol. Assess. 11 (1), 3–13. https://doi.org/10.1037/1040-3590.11.1.3.
Cope, L.M., Ermer, E., Gaudet, L.M., et al., 2014. Abnormal brain structure in youth who commit homicide. NeuroImage Clin. 4, 800–807. https://doi.org/10.1016/j.nicl.2014.05.002.
Cox, J., Edens, J.F., Magyar, M.S., Lilienfeld, S.O., Douglas, K.S., Poythress, N.G., 2013. Using the psychopathic personality inventory to identify subtypes of antisocial personality disorder. J. Crim. Justice 41 (2), 125–134. https://doi.org/10.1016/j.jcrimjus.2012.12.001.
Crego, C., Widiger, T.A., 2015. Psychopathy and the DSM. J. Pers. 83 (6), 665–677. https://doi.org/10.1111/jopy.12115.
Cutler, J., Wittmann, M., Abdurahman, A., et al., 2020. Ageing disrupts reinforcement learning whilst learning to help others is preserved. bioRxiv. ⟨DOI: 10.1101/2020.12.02.407718⟩.
Dashtestani, H., Cui Jr JDH, J., Gandjbakhche, A., 2019. Application of machine learning techniques in investigating the relationship between neuroimaging dataset measured by functional near infra-red spectroscopy and behavioral dataset in a moral judgment task. Clinical and Translational Neurophotonics 2019. SPIE, pp. 48–54. https://doi.org/10.1117/12.2520453.
De Brito, S.A., Mechelli, A., Wilke, M., et al., 2009. Size matters: increased grey matter in boys with conduct problems and callous–unemotional traits. Brain 132 (4), 843–852. https://doi.org/10.1093/brain/awp011.
De Brito, S.A., Forth, A.E., Baskin-Sommers, A.R., et al., 2021a. Psychopathy. Nat. Rev. Dis. Prim. 7 (1), 1–21. https://doi.org/10.1038/s41572-021-00282-1.
De Brito, S.A., McDonald, D., Camilleri, J.A., Rogers, J.C., 2021b. Cortical and subcortical gray matter volume in psychopathy: a voxel-wise meta-analysis. J. Abnorm. Psychol. 130 (6), 627–640. https://doi.org/10.1037/abn0000698.
Deming, P., Heilicher, M., Koenigs, M., 2022. How reliable are amygdala findings in psychopathy? A systematic review of MRI studies. Neurosci. Biobehav. Rev. 142, 104875 https://doi.org/10.1016/j.neubiorev.2022.104875.
Driessen, J.M.A., van Baar, J.M., Sanfey, A.G., Glennon, J.C., Brazil, I.A., 2021. Moral strategies and psychopathic traits. J. Abnorm. Psychol. 130 (5), 550–561. https://doi.org/10.1037/abn0000675.
Edens, J.F., Marcus, D.K., Lilienfeld, S.O., Poythress Jr., N.G., 2006. Psychopathic, not psychopath: taxometric evidence for the dimensional structure of psychopathy. J. Abnorm. Psychol. 115 (1), 131–144. https://doi.org/10.1037/0021-843X.115.1.131.
Fairchild, G., Hagan, C.C., Walsh, N.D., Passamonti, L., Calder, A.J., Goodyer, I.M., 2013. Brain structure abnormalities in adolescent girls with conduct disorder. J. Child Psychol. Psychiatry 54 (1), 86–95. https://doi.org/10.1111/j.1469-7610.2012.02617.x.
Faraone, S.V., Biederman, J., Keenan, K., Tsuang, M.T., 1991. A family-genetic study of girls with DSM-III attention deficit disorder. Am. J. Psychiatry 148 (1), 112–117. https://doi.org/10.1176/ajp.148.1.112.
de Filippis, R., Carbone, E.A., Gaetano, R., et al., 2019. Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review. Neuropsychiatr. Dis. Treat. 15, 1605–1627. https://doi.org/10.2147/NDT.S202418.
Finger, E.C., Marsh, A.A., Blair, K.S., et al., 2011. Disrupted reinforcement signaling in the orbitofrontal cortex and caudate in youths with conduct disorder or oppositional defiant disorder and a high level of psychopathic traits. Am. J. Psychiatry 168 (2), 152–162. https://doi.org/10.1176/appi.ajp.2010.10010129.
Flach, P., 2012. Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge University Press.
Frick, P.J., White, S.F., 2008. Research review: the importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. J. Child Psychol. Psychiatry 49 (4), 359–375. https://doi.org/10.1111/j.1469-7610.2007.01862.x.
Frick, P.J., Kimonis, E.R., Dandreaux, D.M., Farell, J.M., 2003. The 4 year stability of psychopathic traits in non-referred youth. Behav. Sci. Law 21 (6), 713–736. https://doi.org/10.1002/bsl.568.
Frick, P.J., Stickle, T.R., Dandreaux, D.M., Farrell, J.M., Kimonis, E.R., 2005. Callous–unemotional traits in predicting the severity and stability of conduct problems and delinquency. J. Abnorm. Child Psychol. 33 (4), 471–487. https://doi.org/10.1007/s10648-005-5728-9.
Gong, J., Zhang, X., Wang, M.C., Gao, Y., 2022. Latent profile analysis of psychopathy in chinese female offenders. J. Psychopathol. Behav. Assess. https://doi.org/10.1007/s10862-022-09958-8. Published online February 5,.
Gong, X., Brazil, I.A., Chang, L.J., Sanfey, A.G., 2019. Psychopathic traits are related to diminished guilt aversion and reduced trustworthiness during social decision-making. Sci. Rep. 9 (1), 7307. https://doi.org/10.1038/s41598-019-43727-0.
Gregory, S., Blair, R.J., ffytche, D., et al., 2015. Punishment and psychopathy: a case-control functional MRI investigation of reinforcement learning in violent antisocial personality disordered men. Lancet Psychiatry 2 (2), 153–160. https://doi.org/10.1016/S2215-0366(14)00071-6.
Guitart-Masip, M., Huys, Q.J.M., Fuentemilla, L., Dayan, P., Duzel, E., Dolan, R.J., 2012. Go and no-go learning in reward and punishment: interactions between affect and effect. NeuroImage 62 (1), 154–166. https://doi.org/10.1016/j.neuroimage.2012.04.024.
Gullapalli, A.R., Anderson, N.E., Yerramsetty, R., Harenski, C.L., Kiehl, K.A., 2021. Quantifying the psychopathic stare: automated assessment of head motion is related to antisocial traits in forensic interviews. J. Res. Pers. 92, 104093 https://doi.org/10.1016/j.jrp.2021.104093.
Hare, R.D., Neumann, C.S., 2008. Psychopathy as a clinical and empirical construct. Annu. Rev. Clin. Psychol. 217–246.
Hauser, T.U., Will, G.J., Dubois, M., Dolan, R.J., 2019. Annual research review: developmental computational psychiatry. J. Child Psychol. Psychiatry 60 (4), 412–426. https://doi.org/10.1111/jcpp.12964.
Henning, A., 2017. Machine Learning and Natural Language Methods for Detecting Psychopathy in Textual Data (Electronic Theses and Dissertation). ⟨https://egrove.olemiss.edu/etd/446⟩.
Hervé, H., 2007. Psychopathy across the ages: a history of the hare psychopath. The Psychopath: Theory, Research, and Practice. Routledge.
Hicks, B.M., Markon, K.E., Patrick, C.J., Krueger, R.F., Newman, J.P., 2004. Identifying psychopathy subtypes on the basis of personality structure. Psychol. Assess. 16 (3), 276–288. https://doi.org/10.1037/1040-3590.16.3.276.
Hinshaw, S.P., Lahey, B.B., Hart, E.L., 1993. Issues of taxonomy and comorbidity in the development of conduct disorder. Dev. Psychopathol. 5 (1–2), 31–49. https://doi.org/10.1017/S0954579400004247.
Hitczenko, K., Cowan, H.R., Goldrick, M., Mittal, V.A., 2022. Racial and ethnic biases in computational approaches to psychopathology. Schizophr. Bull. 48 (2), 285–288. https://doi.org/10.1093/schbul/sbab131.
Husain, M., Roiser, J.P., 2018. Neuroscience of apathy and anhedonia: a transdiagnostic approach. Nat. Rev. Neurosci. 19 (8), 470–484. https://doi.org/10.1038/s41583-018-0029-9.
Huys, Q.J.M., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19 (3), 404–413. https://doi.org/10.1038/nn.4238.
Jain, D., Arora, S., Jha, C.K., 2019. Diagnosis of psychopathic personality disorder with speech patterns. In: Luhach, A.K., Jat, D.S., Hawari, K.B.G., Gao, X.Z., Lingras, P. (Eds.), Advanced Informatics for Computing Research. Communications in Computer and Information Science. Springer, pp. 411–421. https://doi.org/10.1007/978-981-15-0108-1_38.
Jurjako, M., Malatesti, L., Brazil, I.A., 2019. Some ethical considerations about the use of biomarkers for the classification of adult antisocial individuals. Int. J. Forensic Ment. Health 18 (3), 228–242. https://doi.org/10.1080/14999013.2018.1485188.
Jurjako, M., Malatesti, L., Brazil, I.A., 2020. Biocognitive classification of antisocial individuals without explanatory reductionism. Perspect. Psychol. Sci. 15 (4), 957–972. https://doi.org/10.1177/1745691620904160.
Kahn, R.E., Frick, P.J., Youngstrom, E., Findling, R.L., Youngstrom, J.K., 2012. The effects of including a callous–unemotional specifier for the diagnosis of conduct disorder. J. Child Psychol. Psychiatry 53 (3), 271–282. https://doi.org/10.1111/j.1469-7610.2011.02463.x.
Kimonis, E.R., Fanti, K., Singh, J.P., 2014. Establishing cut-off scores for the parent-reported inventory of callous-unemotional traits. Arch. Forensic Psychol. 1, 27–48.
Kliem, S., Krieg, Y., Klatt, T., Baier, D., 2022. Dimensional latent structure of callous-unemotional traits in german adolescents: results from taxometric analyses. Res. Child Adolesc. Psychopathol. 50 (6), 771–780. https://doi.org/10.1007/s10802-021-00885-y.
Koh, Joel E.W., Ooi, C.P., Lim-Ashworth, N.S.J., et al., 2022. Automated classification of attention deficit hyperactivity disorder and conduct disorder using entropy features with ECG signals. Comput. Biol. Med. 140, 105120 https://doi.org/10.1016/j.compbiomed.2021.105120.
Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Accessed November 15, 2022. https://www.frontiersin.org/articles/ Front. Syst. Neurosci. 2. https://doi.org/10.3389/neuro.06.004.2008.
Latimer, K., Wilson, P., Kemp, J., et al., 2012. Disruptive behaviour disorders: a systematic review of environmental antenatal and early years risk factors. Child Care Health Dev. 38 (5), 611–628. https://doi.org/10.1111/j.1365-2214.2012.01366.x.
Lilienfeld, S.O., Patrick, C.J., Benning, S.D., Berg, J., Sellbom, M., Edens, J.F., 2012. The role of fearless dominance in psychopathy: confusions, controversies, and clarifications. Pers. Disord. Theory Res. Treat. 3 (3), 327–340. https://doi.org/10.1037/a0026987.
Lockwood, P.L., Klein-Flügge, M.C., 2020. Computational modelling of social cognition and behaviour—a reinforcement learning primer. Soc. Cogn. Affect. Neurosci. https://doi.org/10.1093/scan/nsaa040 (Published online).
Lockwood, P.L., Hamonet, M., Zhang, S.H., et al., 2017a. Prosocial apathy for helping others when effort is required. Nat. Hum. Behav. 1 (7), 1–10. https://doi.org/10.1038/s41562-017-0131.
Lockwood, P.L., Ang, Y.S., Husain, M., Crockett, M.J., 2017b. Individual differences in empathy are associated with apathy-motivation. Sci. Rep. 7 (1), 17293. https://doi.org/10.1038/s41598-017-17415-w.
Lockwood, P.L., Apps, M.A.J., Chang, S.W.C., 2020. Is there a 'social' brain? Implementations and algorithms. Trends Cogn. Sci. 24 (10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011.
Loughman, A., Haslam, N., 2018. Neuroscientific explanations and the stigma of mental disorder: a meta-analytic study. Cogn. Res. Princ. Implic. 3 (1), 43. https://doi.org/10.1186/s41235-018-0136-1.
Lu, F., Zhao, Y., He, Z., et al., 2021. Altered dynamic regional homogeneity in patients with conduct disorder. Neuropsychologia 157, 107865. https://doi.org/10.1016/j.neuropsychologia.2021.107865.
Lykken, D.T., 1957. A study of anxiety in the sociopathic personality. J. Abnorm. Soc. Psychol. 55 (1), 6–10. https://doi.org/10.1037/h0047232.

Lynam, D.R., Miller, J.D., 2012. Fearless dominance and psychopathy: a response to Lilienfeld et al. Pers. Disord. Theory Res. Treat. 3 (3), 341–353. https://doi.org/10.1037/a0028296.
Mahmud, S., Rana, M., Zahir, F.R., Huq, M.R., 2021. Detection of antisocial personality based on social media data. In: Fong, S., Dey, N., Joshi, A. (Eds.), ICT Analysis and Applications. Lecture Notes in Networks and Systems. Springer, pp. 651–659. https://doi.org/10.1007/978-981-15-8354-4_65.
Mars, R.B., Shea, N.J., Kolling, N., Rushworth, M.F.S., 2012. Model-based analyses: promises, pitfalls, and example applications to the study of cognitive control. Q J. Exp. Psychol. 65 (2), 252–267. https://doi.org/10.1080/17470211003668272.
Meeks, S.F., 2020. Evaluating Class Imbalance and Asymmetric Costs Using Machine Learning (Thesis). ⟨https://ttu-ir.tdl.org/handle/2346/86573⟩.
Millon, T., Simonsen, E., Birket-Smith, M., Davis, R.D., 2002. Psychopathy: Antisocial, Criminal, and Violent Behavior. Guilford Press.
Moran, P., 1999. The epidemiology of antisocial personality disorder. Soc. Psychiatry Psychiatr. Epidemiol. 34 (5), 231–242. https://doi.org/10.1007/s001270050138.
Morana, H.C.P., Câmara, F.P., Arboleda-Flórez, J., 2006. Cluster analysis of a forensic population with antisocial personality disorder regarding PCL-R scores: differentiation of two patterns of criminal profiles. Forensic Sci. Int. 164 (2), 98–101. https://doi.org/10.1016/j.forsciint.2005.12.003.
Morse, S.J., 1995. Brain and blame essay. Georget. Law J. 84 (3), 527–550.
Moul, C., Robinson, O.J., Livesey, E.J., 2021. Antisocial learning: using learning window width to model callous-unemotional traits? Comput. Psychiatry 5 (1), 54–59.
Newman, J.P., Kosson, D.S., 1986. Passive avoidance learning in psychopathic and nonpsychopathic offenders. J. Abnorm. Psychol. 95 (3), 252–256. https://doi.org/10.1037/0021-843X.95.3.252.
Newman, J.P., MacCoon, D.G., Vaughn, L.J., Sadeh, N., 2005. Validating a distinction between primary and secondary psychopathy with measures of Gray's BIS and BAS constructs. J. Abnorm. Psychol. 114 (2), 319–323. https://doi.org/10.1037/0021-843X.114.2.319.
Nock, M.K., Kazdin, A.E., Hiripi, E., Kessler, R.C., 2006. Prevalence, subtypes, and correlates of DSM-IV conduct disorder in the National Comorbidity Survey Replication. Psychol. Med 36 (5), 699–710. https://doi.org/10.1017/S0033291706007082.
O'Brien, B.S., Frick, P.J., 1996. Reward dominance: associations with anxiety, conduct problems, and psychopathy in children. J. Abnorm. Child Psychol. 24 (2), 223–240. https://doi.org/10.1007/BF01441486.
Oba, T., Katahira, K., Ohira, H., 2019. The effect of reduced learning ability on avoidance in psychopathy: a computational approach. Accessed April 27, 2022. https://www.frontiersin.org/article/ Front. Psychol. 10. https://doi.org/10.3389/fpsyg.2019.02432.
Ogloff, J.R.P., 2006. Psychopathy/antisocial personality disorder conundrum. Aust. N. Z. J. Psychiatry 40 (6–7), 519–528. https://doi.org/10.1080/j.1440-1614.2006.01834.x.
Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G., 2015. Contextual modulation of value signals in reward and punishment learning. Nat. Commun. 6 (1), 8096. https://doi.org/10.1038/ncomms9096.
Palminteri, S., Kilford, E.J., Coricelli, G., Blakemore, S.J., 2016. The computational development of reinforcement learning during adolescence. PLOS Comput. Biol. 12 (6), e1004953 https://doi.org/10.1371/journal.pcbi.1004953.
Pauli, R., Kohls, G., Tino, P., et al., 2021a. Machine learning classification of conduct disorder with high versus low levels of callous-unemotional traits based on facial emotion recognition abilities. Eur. Child Adolesc. Psychiatry. https://doi.org/10.1007/s00787-021-01893-5.
Pauli, R., Tino, P., Rogers, J.C., et al., 2021b. Positive and negative parenting in conduct disorder with high versus low levels of callous–unemotional traits. Dev. Psychopathol. 33 (3), 980–991. https://doi.org/10.1017/S0954579420000279.
Pauli, R., Brazil, I., Kohls, G., et al., 2022. Action initiation and punishment learning differ from childhood to adolescence while reward learning remains stable. ⟨DOI: 10.1101/2022.05.05.490578⟩.
Pearce, M., 2015. Classifying Psychopathy Patients Using Machine Learning Methods on Magnetic Resonance Imaging (MRI) Data, p. 7.
Pike, A.C., Robinson, O.J., 2022. Reinforcement learning in patients with mood and anxiety disorders vs control individuals: a systematic review and meta-analysis. JAMA Psychiatry 79 (4), 313–322. https://doi.org/10.1001/jamapsychiatry.2022.0051.
Prosser, A., Friston, K.J., Bakker, N., Parr, T., 2018. A Bayesian account of psychopathy: a model of lacks remorse and self-aggrandizing. Comput. Psychiatry 2, 92–140. https://doi.org/10.1162/cpsy_a_00016.
Pulcu, E., Browning, M., 2017. Using computational psychiatry to rule out the hidden causes of depression. JAMA Psychiatry 74 (8), 777–778. https://doi.org/10.1001/jamapsychiatry.2017.1500.
Raab, H.A., Hartley, C.A., 2020. Adolescents exhibit reduced Pavlovian biases on instrumental learning. Sci. Rep. 10 (1), 15770. https://doi.org/10.1038/s41598-020-72628-w.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II, second ed. Appleton-Century-Crofts, pp. 64–99.
Research Domain Criteria (RDoC), 2022. National Institute of Mental Health (NIMH). ⟨https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc⟩. (Accessed 15 November 2022).
Rogers, R., Rogstad, J.E., 2010. Psychopathy and APD in non-forensic patients: improved predictions or disparities in cut scores? J. Psychopathol. Behav. Assess. 32 (3), 353–362. https://doi.org/10.1007/s10862-009-9175-8.
Romeo, R., Knapp, M., Scott, S., 2006. Economic cost of severe antisocial behaviour in children - and who pays it. Br. J. Psychiatry 188 (6), 547–553. https://doi.org/10.1192/bjp.bp.104.007625.
Rowe, R., Maughan, B., Moran, P., Ford, T., Briskman, J., Goodman, R., 2010. The role of callous and unemotional traits in the diagnosis of conduct disorder. J. Child Psychol. Psychiatry 51 (6), 688–695. https://doi.org/10.1111/j.1469-7610.2009.02199.x.
Sato, J.R., de Oliveira-Souza, R., Thomaz, C.E., et al., 2011. Identification of psychopathic individuals using pattern classification of MRI images. Soc. Neurosci. 6 (5–6), 627–639. https://doi.org/10.1080/17470919.2011.562687.
Schorr, M.T., Quadors dos Santos, B.T.M., Feiten, J.G., et al., 2021. Association between childhood trauma, parental bonding and antisocial personality disorder in adulthood: a machine learning approach. Psychiatry Res. 304, 114082 https://doi.org/10.1016/j.psychres.2021.114082.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275 (5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593.
Sebastian, C.L., De Brito, S.A., McCrory, E.J., et al., 2016. Grey matter volumes in children with conduct problems and varying levels of callous-unemotional traits. J. Abnorm. Child Psychol. 44 (4), 639–649. https://doi.org/10.1007/s10802-015-0073-0.
Sellbom, M., Drislane, L.E., 2021. The classification of psychopathy. Aggress. Violent Behav. 59, 101473 https://doi.org/10.1016/j.avb.2020.101473.
Sethi, A., Voon, V., Critchley, H.D., Cercignani, M., Harrison, N.A., 2018. A neurocomputational account of reward and novelty processing and effects of psychostimulants in attention deficit hyperactivity disorder. Brain 141 (5), 1545–1557. https://doi.org/10.1093/brain/awy048.
Shipley, S., Arrigo, B.A., 2001. The confusion over psychopathy (II): implications for forensic (correctional) practice. Int. J. Offender Ther. Comp. Criminol. 45 (4), 407–420. https://doi.org/10.1177/0306624X01454002.
Steele, V.R., Rao, V., Calhoun, V.D., Kiehl, K.A., 2017. Machine learning of structural magnetic resonance imaging predicts psychopathic traits in adolescent offenders. NeuroImage 145, 265–273. https://doi.org/10.1016/j.neuroimage.2015.12.013.
Sterzer, P., Stadler, C., Poustka, F., Kleinschmidt, A., 2007. A structural neural deficit in adolescents with conduct disorder and its association with lack of empathy. NeuroImage 37 (1), 335–342. https://doi.org/10.1016/j.neuroimage.2007.04.043.
Suchting, R., Gowin, J.L., Green, C.E., Walss-Bass, C., Lane, S.D., 2018. Genetic and psychosocial predictors of aggression: variable selection and model building with component-wise gradient boosting. Accessed June 20, 2022. https://www.frontiersin.org/article/ Front. Behav. Neurosci. 12. https://doi.org/10.3389/fnbeh.2018.00089.
Sumner, C., Byers, A., Boochever, R., Park, G.J., 2012. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In: Proceedings of the 11th International Conference on Machine Learning and Applications. Vol. 2, pp. 386–393. ⟨DOI: 10.1109/ICMLA.2012.218⟩.
Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning, Second Edition: An Introduction. MIT Press.
Swogger, M.T., Kosson, D.S., 2007. Identifying subtypes of criminal psychopaths: a replication and extension. Crim. Justice Behav. 34 (8), 953–970. https://doi.org/10.1177/0093854807300758.
Tor, H.T., Ooi, C.P., Lim-Ashworth, N.S., et al., 2021. Automated detection of conduct disorder and attention deficit hyperactivity disorder using decomposition and nonlinear techniques with EEG signals. Comput. Methods Prog. Biomed. 200, 105941 https://doi.org/10.1016/j.cmpb.2021.105941.
Vassileva, J., Kosson, D.S., Abramowitz, C., Conrod, P., 2005. Psychopathy versus psychopathies in classifying criminal offenders. Leg. Criminol. Psychol. 10 (1), 27–43. https://doi.org/10.1348/135532504X15376.
Viding, E., Blair, R.J.R., Moffitt, T.E., Plomin, R., 2005. Evidence for substantial genetic risk for psychopathy in 7-year-olds. J. Child Psychol. Psychiatry 46 (6), 592–597. https://doi.org/10.1111/j.1469-7610.2004.00393.x.
Vieira, J.B., Almeida, P.R., Ferreira-Santos, F., Barbosa, F., Marques-Teixeira, J., Marsh, A.A., 2014. Distinct neural activation patterns underlie economic decisions in high and low psychopathy scorers. Soc. Cogn. Affect. Neurosci. 9 (8), 1099–1107. https://doi.org/10.1093/scan/nst093.
Wald, R., Khoshgoftaar, T.M., Napolitano, A., Sumner, C., 2012. Using Twitter content to predict psychopathy. In: Proceedings of the 11th International Conference on Machine Learning and Applications. Vol. 2, pp. 394–401. ⟨DOI: 10.1109/ICMLA.2012.228⟩.
White, S.F., Pope, K., Sinclair, S., et al., 2013. Disrupted expected value and prediction error signaling in youths with disruptive behavior disorders during a passive avoidance task. Am. J. Psychiatry 170 (3), 315–323. https://doi.org/10.1176/appi.ajp.2012.12060840.
White, S.F., Tyler, P.M., Erway, A.K., et al., 2016. Dysfunctional representation of expected value is associated with reinforcement-based decision-making deficits in adolescents with conduct problems. J. Child Psychol. Psychiatry 57 (8), 938–946. https://doi.org/10.1111/jcpp.12557.
, 2018World Health Organization, 2018. ICD-11: International Classification of Diseases 11th Revision: The Global Standard for Diagnostic Health Information. 11th ed. World Health Organization.
Yu, H., Lu, C., Gao, X., et al., 2022. Explaining individual differences in advantageous inequity aversion by social-affective trait dimensions and family environment. Soc. Psychol. Pers. Sci. 13 (2), 626–637. https://doi.org/10.1177/19485506211027794.
Zhang, J., Liu, W., Zhang, J., et al., 2018. Distinguishing adolescents with conduct disorder from typically developing youngsters based on pattern classification of brain structural MRI. Accessed June 22, 2022. https://www.frontiersin.org/article/ Front. Hum. Neurosci. 12. https://doi.org/10.3389/fnhum.2018.00152.

Zhang, J., Cao, W., Wang, M., Wang, N., Yao, S., Huang, B., 2019. Multivoxel pattern analysis of structural MRI in children and adolescents with conduct disorder. Brain Imaging Behav. 13 (5), 1273–1280. https://doi.org/10.1007/s11682-018-9953-6.

Zhang, J., Li, X., Li, Y., et al., 2020a. Three dimensional convolutional neural network-based classification of conduct disorder with structural MRI. Brain Imaging Behav. 14 (6), 2333–2340. https://doi.org/10.1007/s11682-019-00186-5.

Zhang, J., Liu, Y., Luo, R., et al., 2020b. Classification of pure conduct disorder from healthy controls based on indices of brain networks during resting state. Med. Biol. Eng. Comput. 58 (9), 2071–2082. https://doi.org/10.1007/s11517-020-02215-8.

Ziegler, S., Pedersen, M.L., Mowinckel, A.M., Biele, G., 2016. Modelling ADHD: a review of ADHD theories through their predictions for computational models of decision-making and reinforcement learning. Neurosci. Biobehav. Rev. 71, 633–656. https://doi.org/10.1016/j.neubiorev.2016.09.002.