Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/121151

Psychopathy-related traits and the use of reward and social information: a computational approach

Inti A. Brazil[1,2]*, Laurence T. Hunt[3,4], Berend H. Bulten[2], Roy P. C. Kessels[1,5], Ellen R. A. de Bruijn[6] and Rogier B. Mars[1,7,8]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Netherlands
[2] Pompestichting, Nijmegen, Netherlands
[3] Wellcome Trust Centre for Neuroimaging, University College London, London, UK
[4] Sobell Department of Motor Neuroscience, University College London, London, UK
[5] Department of Medical Psychology and Geriatrics, Radboud University Nijmegen Medical Centre, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands
[6] Department of Clinical, Health, and Neuropsychology, Leiden Institute for Brain and Cognition, Leiden University, Leiden, Netherlands
[7] Department of Experimental Psychology, University of Oxford, Oxford, UK
[8] Oxford Centre for Functional MRI of the Brain, University of Oxford, John Radcliffe Hospital, Oxford, UK

Psychopathy is often linked to disturbed reinforcement-guided adaptation of behavior in both clinical and non-clinical populations. Recent work suggests that these disturbances might be due to a deficit in *actively using* information to guide changes in behavior. However, how much information is actually used to guide behavior is difficult to observe directly. Therefore, we used a computational model to estimate the use of information during learning. Thirty-six female subjects were recruited based on their total scores on the Psychopathic Personality Inventory (PPI), a self-report psychopathy list, and performed a task involving simultaneous learning of reward-based and social information. A Bayesian reinforcement-learning model was used to parameterize the use of each source of information during learning. Subsequently, we used the subscales of the PPI to assess psychopathy-related traits, and the traits that were strongly related to the model's parameters were isolated through a formal variable selection procedure. Finally, we assessed how these covaried with model parameters. We succeeded in isolating key personality traits believed to be relevant for psychopathy that can be related to model-based descriptions of subject behavior. Use of reward-history information was negatively related to levels of trait anxiety and fearlessness, whereas use of social advice decreased as the perceived ability to manipulate others and lack of anxiety increased. These results corroborate previous findings suggesting that sub-optimal use of different types of information might be implicated in psychopathy. They also further highlight the importance of considering the potential of computational modeling to understand the role of latent variables, such as the weight people give to various sources of information during goal-directed behavior, when conducting research on psychopathy-related traits and in the field of forensic psychiatry.

Keywords: psychopathy, psychopathic traits, personality traits, individual differences, reinforcement learning, social learning, associative learning, computational modeling

INTRODUCTION

Adults and children with psychopathic tendencies typically show reduced affective-interpersonal functioning, often accompanied by an antisocial lifestyle (Hare et al., 1991; Viding and Larsson, 2007; Sadeh and Verona, 2008; Verona et al., 2012). Research from our own and other labs has shown that offenders with high levels of psychopathic tendencies exhibit deficiencies in associative learning based on reward and punishment (Newman and Kosson, 1986; Budhani et al., 2006; von Borries et al., 2010). It has also been advocated that these deficiencies might lead to impaired associative learning based on social information, resulting in antisocial behavior and a lack of morality (Blair and Cipolotti, 2000; Blair, 2007; Brazil et al., 2011). This claim is also in line with findings in healthy individuals showing that associative learning of

reward and social values follow the same mechanistic principles in the brain, albeit via separable neural substrates (Behrens et al., 2008, 2009).

Results obtained in our lab indicate that psychopathy seems to be related to a reduced ability to actively use information signaling that a change in current behavior is required in order to perform optimally (von Borries et al., 2010; Brazil et al., 2013). To date, however, there has been no direct quantification of how social and reward information is used during associative learning. One reason is that the mainstream experimental approaches in psychiatry do not allow the direct quantification of how much information is used to adapt behavior (see also Montague et al., 2012). However, this limitation can be overcome by incorporating computational modeling of behavior and known neurobiology in understanding

psychiatric conditions (Huys et al., 2011; Maia and Frank, 2011; Buckholtz and Meyer-Lindenberg, 2012). Computational models of associative learning have proven to be increasingly helpful in explaining pathological behavior in neurological disorders like Parkinson's disease (Frank et al., 2004), but also in psychiatric disorders such as schizophrenia (Braver et al., 1999; Fletcher and Frith, 2008) and addiction (Redish et al., 2008). In these conditions, key model parameters can be related to specific aspects of these patients' impaired behavior (Frank et al., 2004) or neurobiology (Corlett et al., 2007), thus allowing the quantification of latent processes that are characteristic of these conditions (i.e., computational phenotypes) (Montague et al., 2012). However, this model-based approach has been notably scarce thus far in research into personality constructs with a less clear conceptual and neurocognitive background such as antisocial personality disorder and psychopathy (Blair, 2005; King-Casas et al., 2008).

There is an on-going debate about the conceptualization of psychopathy (see e.g., Lilienfeld et al., 2012; Miller and Lynam, 2012). Some scholars argue that psychopathy should be defined and assessed in terms of malicious characteristics (e.g., Hare, 2003; Neumann et al., 2012), while others believe that the definition should be broader to also include certain adaptive personality traits (Lilienfeld and Andrews, 1996; Patrick et al., 2009) and there is evidence supporting each approach. Lilienfeld and Andrews (1996) created a questionnaire assessing individual variations in eight common personality traits believed to be strongly related to key adaptive and maladaptive features of psychopathy. Further research suggests that the heightened presence of four of these personality traits may capture part of the aberrant interpersonal-affective personality characteristics and cognitive processing style typical to psychopathy relative to more generic antisocial (i.e. externalizing) personality profiles (see e.g., Poythress et al., 1998; Sadeh and Verona, 2008). The suggestion is that the typical traits are a lack of fear, reduced anxiety, guiltlessness/carelessness/lack of affiliative behavior, and social dominance/manipulative interpersonal style. However, there are very few studies directly relating individual differences in these traits to aspects of psychopathic personality profiles in a quantitative manner (see White et al., 2013).

The main goals of the present study were to use computational modeling to provide the very first direct quantification of the amount of information used to determine behavior during associative learning and to specify which psychopathy-related personality traits are linked to problems in using both social and non-social information. We reasoned that if the diminished *use* of information is a computational phenotype pertaining to psychopathy (relative to generic antisociality), it should also be present among the general population and be related to four personality traits argued to capture aspects of the affective-interpersonal dysfunctions linked to psychopathy and not to the other traits predominantly linked to generic antisociality. To achieve this we sampled a population with varying degrees of common personality traits linked to psychopathy (Lilienfeld and Andrews, 1996; but see Neumann et al., 2012). We then quantified the use of reward history and social advice information to guide behavior in an established reinforcement learning paradigm in which participants have to combine information from both

sources to make optimal choices (Behrens et al., 2008) and used a variable selection method to identify the psychopathy-related traits with the most explanatory power.

METHODS

MEASURE OF PSYCHOPATHY-RELATED TRAITS

Traits were assessed the Dutch translation of the Psychopathic Personality Inventory (PPI) [for more information see (Jelicic et al., 2004)], a self-report questionnaire used to index the presence of traits related to psychopathy in non-clinical samples (Sellbom et al., 2005). Higher scores correspond to higher impact of these traits on personality. The PPI consists of 187 items that are scored on a 4-point Likert scale. Each item loads on one of eight subscales, each subscale representing a different personality trait. The scales are Stress Immunity (displays reduced anxiety), Social Potency (is able to charm and manipulate others/is socially dominant), Fearlessness (lacks fear of harmful consequences), Machiavellian Egocentricity (is self-centered), Blame Externalization (blames others), Carefree Non-planfulness (lacks forethought), Impulsive Non-conformity (is reckless and unconventional) and Coldheartedness (is callous, guiltless).

PARTICIPANT RECRUITMENT

A large pool of potential participants was created through advertisements on a university website and on a national news website with a link to a digital version of the PPI ($N = 485$; 160 males and 325 females). The internal consistency of the subscales was acceptable (Chronbach's alpha = 0.71). Total PPI scores did not differ between males ($N = 160$, Mean = 343, $SD = 39.9$) and females (Mean = 350, $SD = 38$), indicating that scores were distributed equally between genders. Subsequently, total PPI scores were divided in quartiles, and participants were invited based on their scores. Participants from all quartiles (thus, from the entire range of PPI total scores) were invited to take part in the experimental session, but the top and bottom quartiles were over-sampled in order to enhance the presence of extreme scores on both sides of the distribution (Bernat et al., 2011). The experimental sample initially consisted of a single, mixed-gender group of 40 individuals. Unfortunately, only 4 males were willing to participate leading to a strong gender imbalance within the group. Therefore, the male subjects were excluded from further analyses and the final sample consisted of 36 females (for PPI scores see **Table 1**), from which 22 (61%) belonged to the top and bottom quartiles of the selection pool and 14 to the 2nd and the 3rd quartile (39%).

All participants received either course credits or a financial compensation and gave written informed consent. The study was approved by the local ethics committee of the Faculty of Social Sciences at the Radboud University in Nijmegen.

EXPERIMENTAL TASK

Completed 290 trials of a decision-making task in which they had to learn about the probability of receiving reward on two options (blue and green rectangles, **Figure 1**) (Behrens et al., 2008). Subjects repeatedly chose between the two rectangles in order to accumulate points. The number of points available (a random number between 1 and 100) was shown in the center of

Table 1 | Mean total PPI score and subscale scores for the experimental sample (n = 36).

Variable	Mean (*SD*)
Age	22.8 (6.4)
Total PPI score	336 (47.8)
Stress immunity	28.9 (5.4)
Social potency	56.6 (12.8)
Fearlessness	41.3 (10.0)
Coldheartedness	46.4 (7.3)
Blame externalization	31.8 (6.9)
Carefree non-planfulness	40.4 (5.7)
Machavellian egocentricity	54.6 (12.0)
Impulsive non-conformity	33.6 (6.3)

each rectangle; this number was added to the subject's score if the option was chosen and rewarded on that trial. Either blue or green could be correct on each trial, but the probability of the two colors being correct was not equal ($p_{blue} = 1 - p_{green}$). The chance of each color being correct could be inferred based upon the recent outcome history, but was subject to reversals during the course of the experiment (see below). However, the reward magnitudes available were independent of the probabilities of each color being correct; thus, as a result of the difference in reward magnitudes associated with the blue and green options, subjects would sometimes choose to pick the less likely color if it was associated with a higher reward. Subjects saw a red bar onscreen, whose length depicted their current score; they aimed to reach a silver target to win €5, or a gold target to win €7.50.

Subjects simultaneously learnt about the reliability of advice from a social partner. On each trial, subjects received advice (red box around choice in **Figure 1**) about which rectangle to choose from a "human partner" (the experimenter), supposedly playing with them (in reality, the advice was computer-generated). The experimenter sat on the other side of a custom-made shield that divided the room, preventing any visual contact between the participant and the experimenter. Prior to the experiment, both "players" went through the instructions together. The partner's advice constituted what we refer to as the "social information" or "social advice" in the results. The partner's advice was predetermined prior to the experiment (and was, by design, uncorrelated with the reward history-based probability). A cover story was provided such that the partner might be incentivized to give either helpful or unhelpful advice in the experiment, and that this might change during the course of the experiment. In essence, participants were told that during the task the confederate's advice could be either correct or incorrect, but that the confederate was executing a different task and his advices were generated based on this task. Participants saw a demonstration of the task executed by the confederate when receiving the task instructions and they were told that the confederate held no knowledge of the participant's choices, nor whether green or blue was correct. That is, the confederate would provide advice and the computer (which were visibly connected through a cross-over network cable) would map this advice to the appropriate color [for further details see (Behrens et al., 2008)]. Irrespective of whether the advice was

trustworthy or untrustworthy, the subject could exploit the advice to gain further information about which of the two options was the best choice on each trial. After the subject had responded (indicated by the gray box around the choice in **Figure 1**), the correct answer was revealed in the center of the screen, and was then replaced by a fixation point before the next trial began.

In summary, subjects had *three independent* sources of information available on each trial to guide their choices—(i) the magnitude of reward available on each option; (ii) the estimated probability of green/blue yielding reward, based on past experience; (iii) the estimated fidelity of the social partner's advice, based on past experience. The true (underlying) probabilities of both (ii) and (iii) were predetermined such that they varied independently of one another, and underwent several reversals during the course of the experiment (Behrens et al., 2008). This meant that subjects had to continually monitor and learn about each source of information throughout the experiment, and also that each source of information had unique explanatory power in explaining variation in choice behavior. Our key question focused on the degree to which subjects used (ii) and (iii) to guide their choices—a feature of their behavior that can be captured formally with a computational model.

MODELING

We fit a behavioral model to estimate the *influence* of each source of information on each subject's behavior (see mathematical description below). Based on behavioral and neuroimaging results from a previous study (Behrens et al., 2008), the model assumes that subjects use Bayesian reinforcement learning (RL) (Behrens et al., 2007) to track both the probability of green/blue being correct and the probability of receiving truthful advice, and then use this information to guide their behavior. The details of this Bayesian RL model are described in a previous paper (Behrens et al., 2007), and the resulting probabilities are shown in **Figures 1B,C**. The key feature of Bayesian RL is that it allows for a learning rate that *varies* depending upon the current stability or volatility of the environment (Yu and Dayan, 2005; Behrens et al., 2007). To capture the *extent* to which each subject used each source of information in guiding their choices, we fit a model that contains two parameters, $\gamma_{reward\ history}$ and γ_{social}, which have analogous functions for reward history and social information, respectively; importantly, these parameters are independent of the rate at which information is *learnt* in the task (which varies through the task via the RL model, and is not fit as a free parameter). The mathematical role of these parameters is described in equations 1 and 2 in section Mathematical model description, below. Intuitively, however, their role can be thought of as controlling the extent to which a given source of information influenced subject choices, as shown in **Figure 2**. If γ is high for a given source of information, then it means that the objective probability associated with that source of information is *amplified*, i.e., pushed more toward 1 if it is greater than 0.5, and more toward 0 if it is less than 0.5 (e.g., the steepest line in **Figure 2A**). Conversely, if γ is low, the objective probability is pulled toward 0.5, and so has less influence (e.g., the shallowest line in **Figure 2A**). We estimated these parameters (and a further temperature parameter
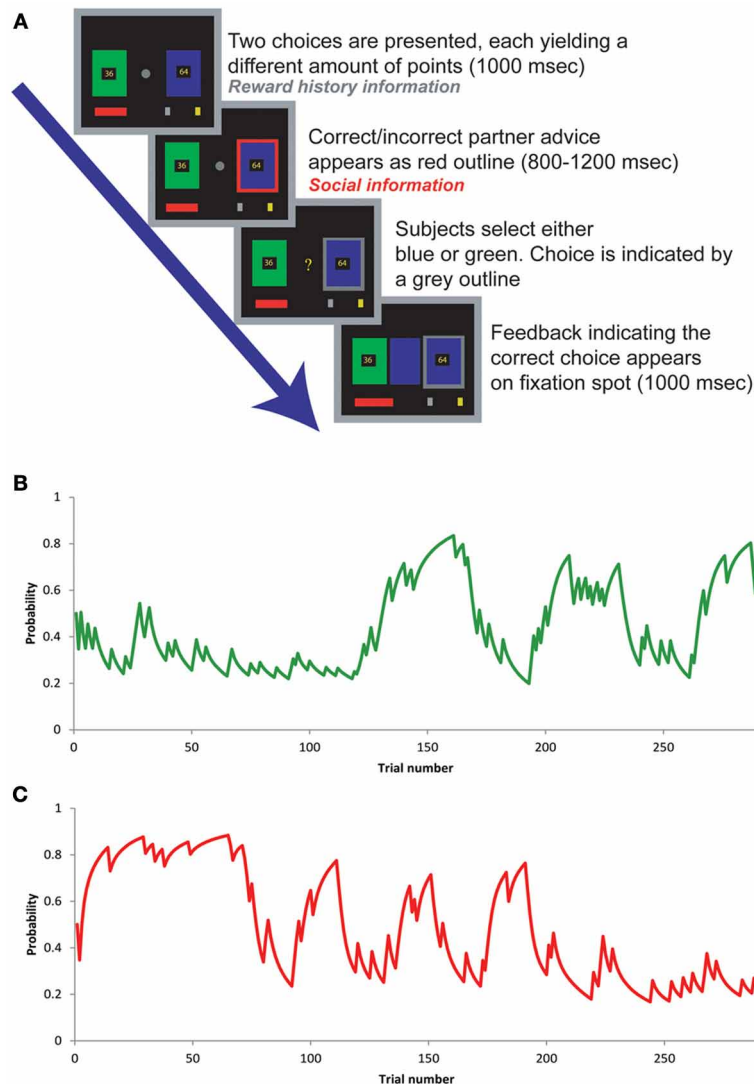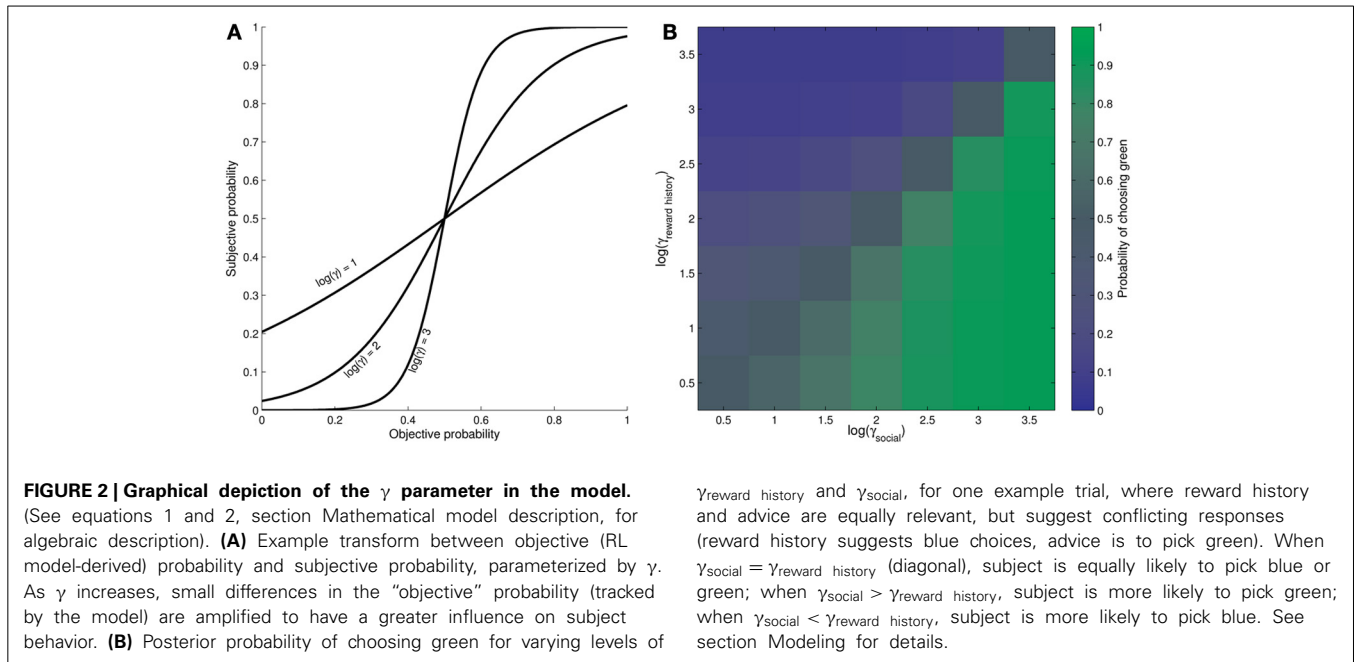
FIGURE 1 | (A) Sequence of events and their timings during the experiment. **(B)** Probability of reward from choosing green card through the experiment. The line shows the probability estimated by the Bayesian reinforcement learning model. **(C)** The figure shows the model-derived probability of the confederate providing the correct answer through the experiment. Note that the model learns *independently* about both social and reward history information at the time feedback is received.

β, capturing choice stochasticity) separately for each subject (see below), in order to investigate cross-subject variability in their expression.

The magnitudes of $\gamma_{\text{reward history}}$ and γ_{social} then become important when we *combine* the sources of information to obtain an *overall probability of selecting green* on each trial. This is illustrated in **Figure 2B**, where we show the effect of varying the two parameters on the eventual probability of the subject wanting to select green for an example trial. In this trial, there is a 0.3 probability of green being rewarded given the recent reward history. However, the confederate has advised green, and there is a 0.7 probability that the confederate will give good advice. Hence, these two sources of information would cancel one another out— but *only* if the subject uses each source of information equally

(i.e., $\gamma_{\text{reward history}} = \gamma_{\text{social}}$). Conversely, if $\gamma_{\text{social}} > \gamma_{\text{reward history}}$, then the subject will favor the social information and become more likely to pick green (green area in **Figure 2B**), whereas if $\gamma_{\text{reward history}} > \gamma_{\text{social}}$, the subject will become more likely to pick blue (blue area in **Figure 2B**). Note that for simplicity, we have shown an example where the points on green and blue are equal; however, further interactions occur with the number of points available as these vary from trial to trial, and also as the probabilities of social and non-social information fluctuate independently of one another. In particular, subjects with small values of $\gamma_{\text{reward history}}$ and γ_{social} are likely to downweight information relating to the past history of reward/social outcomes, and upweight information relating to current reward magnitudes.

FIGURE 2 | Graphical depiction of the γ parameter in the model. (See equations 1 and 2, section Mathematical model description, for algebraic description). **(A)** Example transform between objective (RL model-derived) probability and subjective probability, parameterized by γ. As γ increases, small differences in the "objective" probability (tracked by the model) are amplified to have a greater influence on subject behavior. **(B)** Posterior probability of choosing green for varying levels of

$\gamma_{\text{reward history}}$ and γ_{social}, for one example trial, where reward history and advice are equally relevant, but suggest conflicting responses (reward history suggests blue choices, advice is to pick green). When $\gamma_{\text{social}} = \gamma_{\text{reward history}}$ (diagonal), subject is equally likely to pick blue or green; when $\gamma_{\text{social}} > \gamma_{\text{reward history}}$, subject is more likely to pick green; when $\gamma_{\text{social}} < \gamma_{\text{reward history}}$, subject is more likely to pick blue. See section Modeling for details.

MATHEMATICAL MODEL DESCRIPTION

The model takes estimates of the probability of receiving good advice ($p_{social,i}$) and the probability of green being rewarded ($p_{green,i}$) at trial i, estimated via a Bayesian reinforcement learning optimized for adapting behavior depending upon the underlying volatility of the environment [see **Figures 1B,C** for graphs of tracked probabilities; for details of probability-tracking problem see (Behrens et al., 2007)]. These probability estimates are converted into *subjective* probabilities using the following transforms:

$$\hat{p}_{\text{social},i} = \frac{1}{1 + e^{-\gamma_{\text{social}}\left(p_{\text{social},i}-0.5\right)}} \quad (1)$$

$$\hat{p}_{\text{green},i} = \frac{1}{1 + e^{-\gamma_{\text{reward history}}\left(p_{\text{green},i}-0.5\right)}} \quad (2)$$

These subjective probabilities are then converted into an overall subjective probability of green yielding reward, q_i:

$$\hat{q}_i = \frac{\hat{p}_{\text{social},i}\,\hat{p}_{\text{green},i}}{\hat{p}_{\text{social},i}\,\hat{p}_{\text{green},i} + (1-\hat{p}_{\text{social},i})(1-\hat{p}_{\text{green},i})} \quad (3)$$

if the partner suggests green on trial i, and

$$\hat{q}_i = \frac{\hat{p}_{\text{social},i}\left(\hat{p}_{\text{green},i}\right)}{\hat{p}_{\text{social},i}\left(1-\hat{p}_{\text{green},i}\right) + (1-\hat{p}_{\text{social},i})\,\hat{p}_{\text{green},i}} \quad (4)$$

if the partner suggests blue.
The overall expected value of each option is then calculated as:

$$V_{\text{green},i} = \hat{q}\,r_{\text{green},i} \quad (5)$$

and

$$V_{\text{blue},i} = (1 - \hat{q})r_{\text{blue},i} \quad (6)$$

where r_{green},i and r_{blue},i are the number of points available on green and blue options, respectively, on trial i. Finally, the probability of choosing the green option at trial i is calculated via a softmax function (O'Doherty et al., 2004):

$$P\left(C_i = \text{green}\right) = \frac{1}{1 + e^{-\beta(V_{\text{green}}-V_{\text{blue}})}} \quad (7)$$

and

$$P\left(C_i = \text{blue}\right) = 1 - P(C_i = \text{green}) \quad (8)$$

where β is an additional, third free parameter that determines the stochasticity of choice behavior.
We then used this model to estimate the log-likelihood of the observed data, at given values of the parameters $\gamma_{\text{reward history}}$, γ_{social}, and β:

$$LL\left(\gamma_{\text{social}}, \gamma_{\text{reward history}}, \beta\right) = \sum_i \log\left[P\left(C_i = c_i | \gamma_{\text{social}},\right.\right.$$
$$\left.\left.\gamma_{\text{reward history}}, \beta\right)\right] \quad (9)$$

where c_i denotes the option chosen by the subject on trial i. We custom-implemented a Bayesian estimation procedure in MATLAB (MathWorks, MA) to obtain the best-fitting parameters γ_{social}, $\gamma_{\text{reward history}}$ and β. Specifically, we performed direct numerical integration over the likelihood function of the observed data given the three free parameters. A grid of all possible parameter values of interest was formed, and we evaluated the likelihood of the data at each point in the grid, and then

used marginalization to calculate the marginal likelihood of each parameter. All parameters were allowed to take values between 0.01 and 10, and the grid for numerical integration was evaluated in log space. This approach was selected because it gave a direct measure of the uncertainty associated with each parameter (i.e., the variance of each parameter's posterior distribution), in order to assess the reliability of model fitting.

RELATING FITTED MODEL PARAMETERS TO VARIATIONS IN TRAITS

The key question addressed here is which psychopathy-related traits are linked to the *between-subject variation* in the degree to which each optimally-tracked source of information is used to guide behavior, which is indexed in the model by the free parameters $\gamma_{reward\ history}$ and γ_{social}. To test this, we conducted two separate optimal scaled variable selections using the CATREG module in SPSS. This was done in order to establish the subscales of the PPI with the highest contributions in explaining the variance of each free parameter. For optimal scaling, all variables were defined on a numeric scale and discretized using a multiplication method, which transforms the variables into z-scores and multiplies them by 10. Two models were created which included all subscales of the PPI and the estimates for $\gamma_{reward\ history}$ and γ_{social}, respectively. Subsequently, variable selection with *lasso* [least absolute shrinkage and selection operator; (Tibshirani, 1996)] regularization was implemented to identify the "optimal" model for each free parameter. The optimal model was taken to be the model with the lowest expected prediction error and thus the highest accuracy given the data. This approach relies on shrinking the sum of the model coefficients by adding penalty terms to the model, resulting in coefficients that represent independent contributions of each variable as well as better model accuracy (Hartmann et al., 2009). For the regularization, the minimum of the standardized sum of squares was set at 0.0 and the maximum at 1.0 with a 0.02 increment in shrinkage at each step. This procedure yields an optimal model, which is the model with the smallest predicted margin of error. The latter was estimated with 0.632 bootstrapping (100 samples) (Efron and Tibshirani, 1993).

One advantage of this selection approach is that it overcomes a lot of the limitations of variable selection when using traditional stepwise regression analyses, such as the need for normality of variables (Hartmann et al., 2009), the related loss of power due to lack of compliance with assumptions, and the need for multiple comparison corrections associated with frequentist testing. After selection of the optimal model for each computational parameter, Pearson correlations were calculated between the scales in each model and the corresponding computational parameter in order to establish whether these covary. The significance of the correlations was tested with a non-parametric bootstrapping procedure (10.000 samples) to determine the confidence interval (CI) of each of the scales resulting from the variable selection procedure. If a correlation is significant its CI should not include the value of exactly 0. Thus, *both* the upper and lower bound of a CI should be either larger or smaller than 0.00. Finally, the oversampling procedure might have led to an atypical/non-normal distribution of the total and the scale scores of the PPI. Although our methodological approach did not rely on classical testing procedures requiring compliance with the assumption of normality, we still conducted Kolmogorov-Smirnov (KS) tests of normality to check whether the distribution of the total and scale scores of the PPI in the experimental sample was normal.

RESULTS

GENERAL TEST OF PERFORMANCE

First, we carried out an initial check to ascertain that participants were learning and were engaged in the task by comparing the amount of points earned at the end of the task with chance level performance. The results showed that the average amount of points earned (Mean $= 10.372$, $SD = 780$) was significantly higher than the amount that could be earned by guessing the correct choice on each trial (Mean $= 7.292$, $SD = 577$; $t_{(35)} = 20.3$, $p < 0.001$), indicating above chance performance and that participants were actively engaged in the task. Next, we also checked that the model provided a robust and reliable description of subject behavior. We found that the model, after parameter fitting, accurately predicted which of the two options subjects would choose on $80.6 \pm 7.2\%$ [mean $\pm$ standard deviation (SD)] of trials, indicating that it provided a robust description of subject behavior. Moreover, the uncertainty of estimated parameters (the SD of the posterior distribution) was relatively small compared to the magnitude/range of the estimated parameters (Mean $\gamma_{reward\ history} = 1.14$, SD range $= 0.21–1.1$; Mean $\gamma_{social} = 2.18$, SD range $= 0.18–1.27$), indicating that parameter fitting was reliable.

VARIABLE SELECTION

Here, we present the results of the two variable selection procedures run after the estimation of $\gamma_{reward\ history}$ and γ_{social}, which are displayed in **Figure 3**. The initial model is depicted at the far right of each panel. The systematic shrinkage of the standardized sum of coefficients forces the coefficients toward zero and for each step the resulting model is depicted to the left of the previous model. In both panels, the dashed vertical line indicates the optimal model. Note that for our purpose of solely identifying variables with the greatest contribution to the computational parameters, the magnitude and significance of the variable coefficients (indexed on the Y-axis) are of less interest and that the results do not warrant statistical significance in subsequent tests. Stress Immunity and Fearlessness were the traits that had the largest contributions to the variability across subjects of $\gamma_{reward\ history}$ (**Figure 3A**). In contrast, the optimal model for γ_{social} included the variable Stress Immunity and Social Potency (**Figure 3B**).

CORRELATIONS

Subsequent correlation analyses yielded significant negative correlations between $\gamma_{reward\ history}$ and Stress Immunity ($r = -0.36$, 95% CI -0.60 to -0.04) and $\gamma_{reward\ history}$ and Fearlessness ($r = -0.34$, 95% CI -0.59 to -0.02). The correlation analyses revealed a negative relationship between γ_{social} and Social Potency ($r = -0.34$, 95% CI -0.59 to -0.06) and for γ_{social} and Stress Immunity ($= -0.32$, 95% CI -0.57 to -0.02). Thus, specific traits were related to different computational parameters

quantifying individual difference in the use of reward and social information (see **Figure 4**).

ADDITIONAL TESTS
Additional correlation analyses
In order to demonstrate that the two computational parameters were uncorrelated and that the traits identified were uniquely

related to either γ_{social} (Range = 1.57–24.2) or $\gamma_{reward\ history}$ (Range = 0.42–12.9), we additionally examined the correlations between (1) γ_{social} and $\gamma_{reward\ history}$, (2) Stress Immunity and Fearlessness with γ_{social} and (3) Social Potency with $\gamma_{reward\ history}$. As expected, the computational parameters were not significantly correlated ($r = 0.11$, 95% CI −0.18 to 0.58). Fearlessness was uncorrelated with γ_{social} ($r = -0.10$, 95% CI −0.44 to 0.33), as
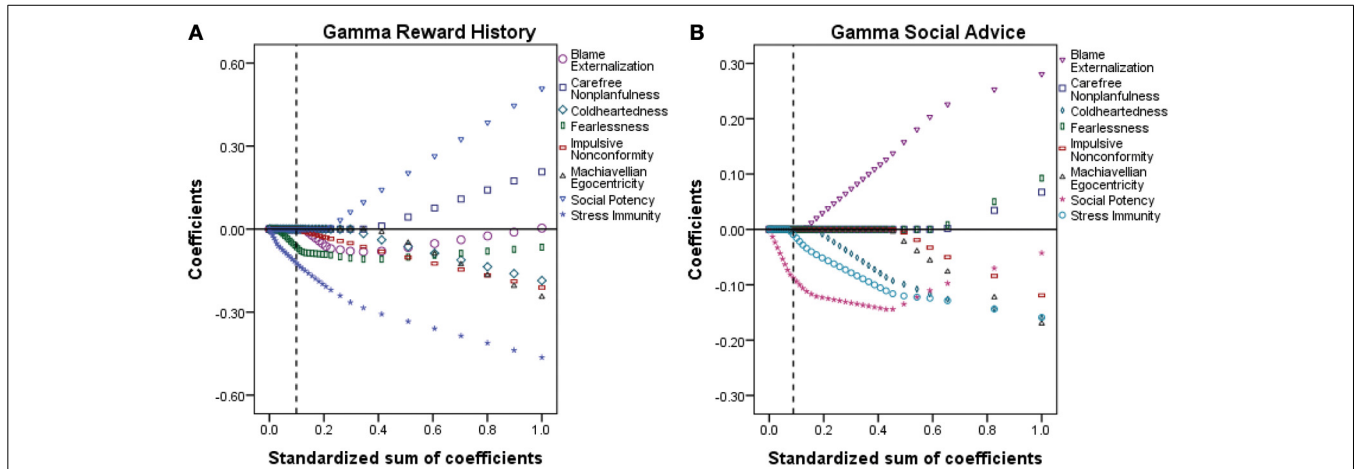
FIGURE 3 | Results of the variable selection procedure for $\gamma_{reward\ history}$ (A) and γ_{social} (B). The maximum standardized sum of coefficients (SSC; x-axis) was set at 1.0, representing 100% of the contribution of the PPI scales to the corresponding γ parameter. Each sub-figure should be read from right

(SSC = 1.0) to left (SSC = 0.0). The variable coefficients (y-axis) are displayed for different stages of shrinkage of the SSC. For each analysis, the variables included in the optimal model (i.e., the model with the lowest expected prediction error) are indicated with the vertical dashed line.
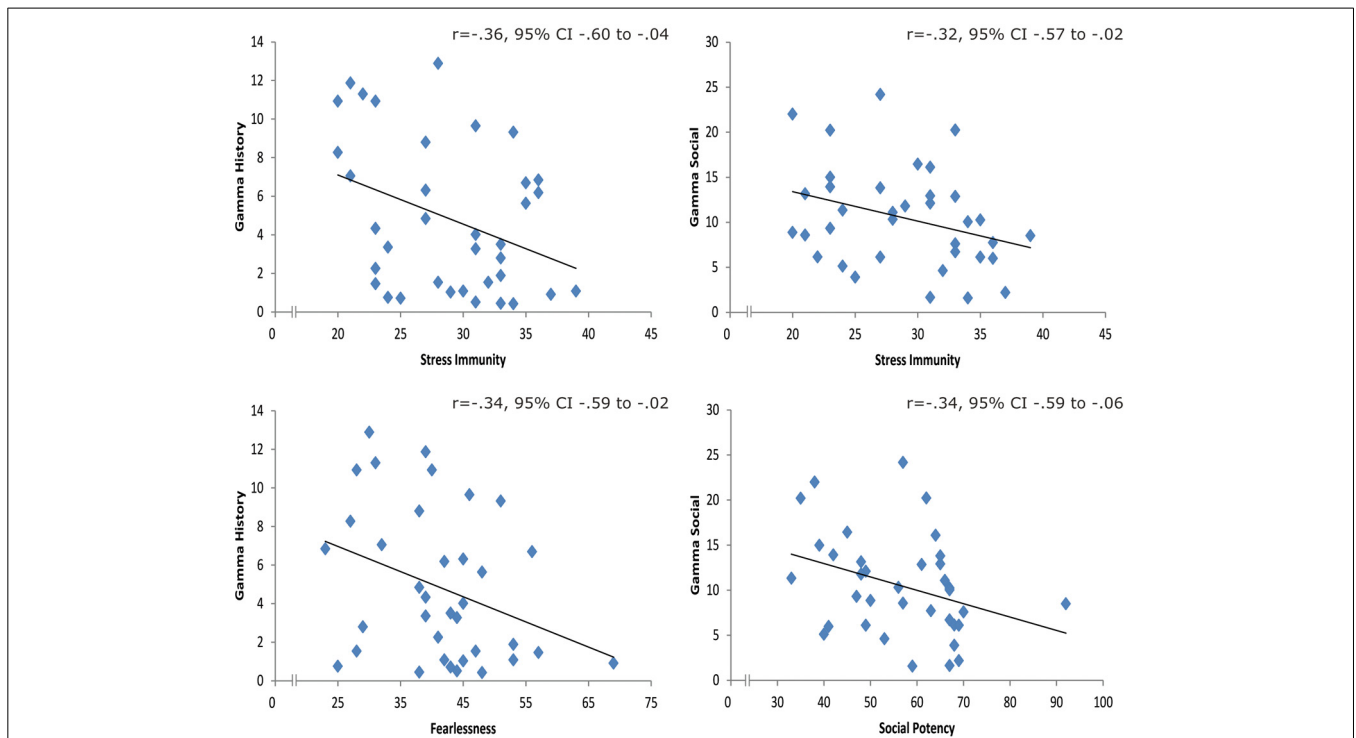

FIGURE 4 | Left: scatterplots for the correlations between $\gamma_{reward\ history}$ and Stress Immunity **(top left)**/Fearlessness **(bottom left)**. **Right:** scatterplots for the correlations between γ_{social} and Stress Immunity **(top right)**/Social Potency **(bottom right)**.

was Social Potency with $\gamma_{reward\ history}$ ($r = -0.07$, 95% CI -0.35 to 0.25). These results indicate contributions of the different traits to the explained variance of the estimated model parameters. The tests of normality showed that the oversampling of the distribution tails in the selection pool ($N = 485$) did not cause the PPI measures in the experimental sample ($n = 36$) to deviate from normality (all KS-Z ≤ 0.95, p's ≥ 0.33).

Comparison with an alternative computational model

Finally, we addressed concerns that our results may be a consequence of a use of a particular model, as opposed to a sensitive measure of the use of social information. We ran a direct comparison of a model that uses the Bayesian probability-tracking scheme and a Rescorla-Wagner learning model that has free parameters for learning rates (social and non-social). The correlation coefficient between γ_{social} for the Bayesian model, and γ_{social} for the fixed learning rate model, was 0.84; the correlation coefficient between $\gamma_{reward\ history}$ for the Bayesian model, and $\gamma_{reward\ history}$ for the fixed learning rate model, was 0.79. Thus, the fit parameters were not heavily influenced by the specific reinforcement learning model used, indicating that the results reported above paper are robust to the precise formulation of the RL model.

However, we elected to use the Bayesian RL model in the analysis above, because comparisons of model evidence vastly favored the Bayesian model. In 32 out of 36 subjects, the Bayesian Information Criterion favored the model with the Bayesian learning rate [paired T-test between BICs: $T_{(35)} = 5.45$, $p < 0.000005$ in favor of Bayesian model]. Similarly, in 25 out of 36 subjects, the Akaike Information Criterion, which has a smaller penalty than BIC for models with more free parameters (such as the fixed learning rate model), still favored the model with the Bayesian learning rate.

DISCUSSION
MAIN FINDINGS
The present study is the first to use formal computational modeling to quantify how information from different sources is used during associative learning in order to provide evidence that variations in personality traits linked to psychopathy are differentially related to diminished use of social and reward information. This was achieved by establishing which specific traits related to psychopathy covary with the ability to actively use social and reward information to guide behavior as indicated by a computational model's parameter fits based on each individual participant's data. In this way, we succeeded in quantifying latent variables that cannot be observed overtly using traditional experimental approaches (Mars et al., 2012), and were able to relate these to personality traits proposed to be associated with core aspects of the construct of psychopathy.

We found that the extent to which participants tended to use reward and social information was related to different personality traits. Traits capturing lack of anxiety (Stress Immunity) and lack of fear (Fearlessness) were negatively correlated with the extent to which previous reward history was used to make decisions. The use of social information was found to have a negative relationship with participants' perceived ability to charm and manipulate others (Social Potency) and lack of anxiety. Importantly, our

effects are selectively associated with personality traits argued to be central to psychopathy, while none of the traits more related to externalizing personality styles were substantially linked to the computational parameters in the present study. In other words, the results suggest that the deficient use of reward and social information during learning could be specific to psychopathic personality styles rather than general antisociality, and also that the deficient implementation of information that seems to be present in male offenders diagnosed with a psychopathic disorder translates to common personality traits linked to psychopathic tendencies in the non-clinical female population.

COMPARISON WITH PREVIOUS WORK
The use of previous reward history was negatively correlated with scores on Stress Immunity and Fearlessness. These findings converge with evidence relating both low anxiety and low fear to disturbed associative learning in clinical psychopathy (Arnett et al., 1993; Birbaumer et al., 2005). Particularly, work by Newman and colleagues has shown that disturbed passive avoidance learning is predominantly found in individuals with psychopathy with low trait anxiety relative to those with high anxiety (Newman et al., 1990; Arnett et al., 1993). Similarly, psychopathic behavior has also repeatedly been linked to reduced fear reactivity in both clinical and non-clinical samples (Patrick et al., 1993; Blair et al., 2002; Benning et al., 2005; Jones et al., 2009) and, importantly, impaired fear-conditioning (Flor et al., 2002; Birbaumer et al., 2005). The central premise here is that aversion to negative outcomes induces a negative affective state such as fear/anxiety, which is in turn associated with the actions/contexts that lead to these negative affective states. With respect to psychopathy, it has been proposed that a low propensity to experience these negative affective states plays a role in the formation of weak associations with events leading to negative outcomes and thus contribute to an impairment in the process of associative learning (Blair, 2005). Our results add support to this notion by pointing out that increased trait fearlessness and lack of anxiety contribute to reduced use of information to guide behavior during associative learning.

One important consideration is that in tasks using behavioral performance as an index for associative learning, these outcome measures not only represent the integrity of the associative process (i.e., the linking sensory events to outcomes) but also the individual's ability to integrate and use relevant sensory information to initiate and execute motor responses/observable behavior (Daunizeau et al., 2010). Thus, covert behavior is the integrated end-result of various processing steps in different domains. Therefore, impaired performance could reflect deficient processing in the sensory domain (e.g., the establishment of associations/learning), or in the motor domain (e.g., execution errors), or maybe a problem in the interaction between the sensory domain and the motor domain (e.g., using learned associations as input to guide motor responses). The present findings indicate that trait fear and anxiety play an important role in the active implementation of available information to guide changes in behavior. This suggests that impairments in associative learning previously found in clinical psychopathy might also be (partly) due to a deficiency in *using* reinforcement information appropriately to drive behavior, which, depending upon the

experimental paradigm used, may ultimately manifest itself as disturbed learning.

The use of information provided by the confederate, i.e., the use of social information history, was found to have a negative relationship with participants' perceived ability to charm and manipulate others (Social Potency) and their level of trait anxiety (Stress Immunity). Social Potency and anxiety encompass behavior relevant for social functioning. High Social Potency is commonly associated with social dominance and one's belief that one is able to successfully manipulate others. We could hypothesise that people who believe that they can manipulate others are more likely to believe that others will try to manipulate them, when *mentalizing* about the likely intentions of the social partner (Behrens et al., 2008; Hampton et al., 2008; Chang et al., 2011). That is, these individuals may be more likely to engage in making inferences about what others may think we believe, i.e., second-order beliefs. A possible explanation for the relationship between lack of anxiety and use of social advice could be that as trait anxiety decreases, individuals experience less anxiety evoked by the potential negative consequences of discarding the confederate's advice. Thus, as individual levels of trait anxiety decrease, not using social advice might be experienced as less aversive, in a way similar to reward-based learning. This prediction would be in line with findings showing that associative learning of social and non-social information follow the same mechanistic principles (Behrens et al., 2008). In sum, our results suggest that reduced anxiety and second-order belief systems might play an important role in explaining social cognition in psychopathy. Future studies should focus on mapping how second-order beliefs are related to general traits relevant to psychopathy in the general community as well as in offenders with a clinical diagnosis of psychopathy.

INTERPRETATIONAL LIMITATIONS

This is one of the first studies that has attempted to link scores on psychopathy-related personality traits with latent variables from a computational model that was fit to each participant's behavior (see also White et al., 2013). This approach has been suggested to have tremendous potential in the study of psychopathology and in psychiatry in general, as it has the potential to be able to disentangle separate aspects of complex multidimensional syndromes (Montague et al., 2012). However, this does not mean that the approach is not without its limitations. Below we suggest some potential improvements and avenues for future studies.

One potential caveat is that in our current model the learning rates for reward and social information were not allowed to vary across subjects. This is due to limitations in the number of trials we would need to reliably estimate more free parameters. Instead, the model used (Behrens et al., 2007) was one that adapts its learning rate dependent upon the current level of volatility in the environment. In the current study, we instead set out to test the hypothesis that the use of different types of information is related to different personality traits that are relevant for psychopathy. The present study included a sample of healthy individuals and previous studies have shown that healthy individuals are able to estimate the volatility of the environment and adapt their learning rate accordingly, and that this behavior is reproduced reliably by our computational model (Behrens et al.,

2008). Future computational studies could be designed to explicitly test the hypothesis that it is use of information rather than (only) learning rate in general that is impaired in offenders diagnosed with psychopathic disorder according to the Psychopathy Checklist-Revised (PCL-R) (Hare, 2003), as suggested by some of our previous findings (von Borries et al., 2010; Brazil et al., 2013).

Another potential limitation of our current study is the size of our group of participants. Although we have used a large sample of participants compared to most computational modeling studies (e.g., Nieuwenhuis et al., 2005; Behrens et al., 2008; Yoshida et al., 2008; Boorman et al., 2009; Mars et al., 2012; Brodersen et al., 2013), some may argue that it is on the lower side in studies in psychological research on personality. We have taken care to ensure the robustness of our effects through the methodology employed, but the size of our sample can still be raised as a criticism despite the fact that our methodology bypasses the need for compliance with the requirements of classical inferencing [for more details on the overlooked issues with various common beliefs about sampling and sample sizes we highly recommend (Friston, 2012, 2013)]. Furthermore, the fact that previous studies using our model found robust results even with much lower subject numbers is therefore quite reassuming (e.g., Behrens et al., 2008; Boorman et al., 2009).

Finally, our experimental sample consisted of female participants and it could be argued that the findings might not extent to the male population. However, previous studies in clinical psychopathy suggesting deficient use of information to adapt behavior included only male participants (Brazil et al., 2009, 2013; von Borries et al., 2010) and as the current results converge with those obtained in male-only samples they support the notion that this particular deficiency in using information to guide behavior does not seem gender-specific. In support of this claim, recent studies on the relationship between psychopathic traits in community samples, empathic responding and moral processing suggest a similar relationship in both males and females (Seara-Cardoso et al., 2012, 2013). Interestingly, Seara-Cardoso et al. (2013) found a negative relationship between these cognitive functions and the interpersonal-affective traits in females. In this study they used a different operationalization of psychopathy (Paulhus et al., 2013) and assessed other aspects of cognitive functioning relative to the present study, but the findings are in line with ours in that they point out that gender might not have an overall impact on the link between psychopathy-related traits and certain aspects of cognition.

CONCLUSIONS

The present study is the first to directly assess the relationship between variations in psychopathy-related personality traits and the amount of information that is used during associative learning of social and reward information. The findings show that the use of both types of information to guide behavior decreases as the presence of personality traits proposed to be related to the interpersonal-affective aspect of psychopathy increases. More specifically, lower trait anxiety and fearlessness were associated with reduced use of one's reinforcement history and an increased perceived ability to manipulate others and reduced anxiety were related to diminished use of social advice. Additionally,

the findings suggest an extension of results obtained in male offenders with clinical psychopathy to the general (female) population by showing that the newly-discovered latent variables are linked to variations in personality traits that are important for the construct of psychopathy. Importantly, however, it still remains to be investigated whether these computational parameters can account for some of the impairments in adaptive behavior found in forensic psychiatric populations with a psychopathic disorder. The results illustrate the potential advantages of employing formal models to discover computational phenotypes in clinical populations (Montague et al., 2012), as well as their usefulness in gaining more insight into the exact personality traits related to the cognitive deficiencies observed in many personality disorders. The present findings might also have implications for treatment aimed at altering behavior, as the success of treatment partly relies on the patient's ability to incorporate and use information from past experience as well as information provided by therapists.

REFERENCES

Arnett, P. A., Howland, E. W., Smith, S. S., and Newman, J. P. (1993). Autonomic responsivity during passive avoidance in incarcerated psychopaths. *Pers. Individ. Dif.* 14, 173–184. doi: 10.1016/0191-8869(93)90187-8

Behrens, T. E. J., Hunt, L. T., and Rushworth, M. F. S. (2009). The computation of social behavior. *Science* 324, 1160–1164. doi: 10.1126/science.1169694

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., and Rushworth, M. F. S. (2008). Associative learning of social value. *Nature* 456, 245–249. doi: 10.1038/nature07538

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221. doi: 10.1038/nn1954

Benning, S. D., Patrick, C. J., and Iacono, W. G. (2005). Psychopathy, startle blink modulation, and electrodermal reactivity in twin men. *Psychophysiology* 42, 753–762. doi: 10.1111/j.1469-8986.2005.00353.x

Bernat, E. M., Nelson, L. D., Steele, V. R., Gehring, W. J., and Patrick, C. J. (2011). Externalizing psychopathology and gain–loss feedback in a simulated gambling task: dissociable components of brain response revealed by time-frequency analysis. *J. Abnorm. Psychol.* 120, 352–364. doi: 10.1037/a0022124

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., et al. (2005). Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. *Arch. Gen. Psychiatry* 62, 799–805. doi: 10.1001/archpsyc.62.7.799

Blair, R.J., and Cipolotti, L. (2000). Impaired social response reversal. a case of "acquired sociopathy." *Brain* 123, 1122–1141. doi: 10.1093/brain/123.6.1122

Blair, R. J. R. (2005). Applying a cognitive neuroscience perspective to the disorder of psychopathy. *Dev. Psychopathol.* 17, 865–891. doi: 10.1017/S0954579405050418

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends Cogn. Sci.* 11, 387–392. doi: 10.1016/j.tics.2007.07.003

Blair, R. J. R., Mitchell, D. G., Richell, R. A., Kelly, S., Leonard, A., Newman, C., et al. (2002). Turning a deaf ear to fear: impaired recognition of vocal affect in psychopathic individuals. *J. Abnorm. Psychol.* 111, 682–686. doi: 10.1037/0021-843X.111.4.682

Boorman, E. D., Behrens, T. E., Woolrich, M. W., and Rushworth, M. F. (2009). How green is the grass on the other side? frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* 62, 733–743. doi: 10.1016/j.neuron.2009.05.014

Braver, T. S., Barch, D. M., and Cohen, J. D. (1999). Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biol. Psychiatry* 46, 312–328. doi: 10.1016/S0006-3223(99)00116-X

Brazil, I. A., de Bruijn, E. R. A., Bulten, B. H., von Borries, A. K. L., van Lankveld, J. J. D. M., Buitelaar, J. K., et al. (2009). Early and late components of error monitoring in violent offenders with psychopathy. *Biol. Psychiatry* 65, 137–143. doi: 10.1016/j.biopsych.2008.08.011

Brazil, I. A., Maes, J. H. R., Scheper, I., Bulten, B. H., Kessels, R. P., Verkes, R. J., et al. (2013). Reversal deficits in psychopathy in explicit but not implicit learning conditions. *J. Psychiatry Neurosci.* 38, e13–e20. doi: 10.1503/jpn.120152

Brazil, I. A., Mars, R. B., Bulten, B. H., Buitelaar, J. K., Verkes, R. J., and De Bruijn, E. R.. (2011). A neurophysiological dissociation between monitoring one's own and others' actions in psychopathy. *Biol. Psychiatry* 69, 693–699. doi: 10.1016/j.biopsych.2010.11.013

Brodersen, K. H., Daunizeau, J., Mathys, C., Chumbley, J. R., Buhmann, J. M., and Stephan, K. E. (2013). Variational bayesian mixed-effects inference for classification studies. *Neuroimage* 76, 345–361. doi: 10.1016/j.neuroimage.2013.03.008

Buckholtz, J. W., and Meyer-Lindenberg, A. (2012). Psychopathology and the human connectome: toward a transdiagnostic model of risk for mental illness. *Neuron* 74, 990–1004. doi: 10.1016/j.neuron.2012.06.002

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

Chang, L. J., Smith, A., Dufwenberg, M., and Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* 70, 560–572. doi: 10.1016/j.neuron.2011.02.056

Corlett, P. R., Murray, G. K., Honey, G. D., Aitken, M. R. F., Shanks, D. R., Robbins, T. W., et al. (2007). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain* 130, 2387–2400. doi: 10.1093/brain/awm173

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., and Friston, K. J. (2010). Observing the observer (i): meta-bayesian models of learning and decision-making. *PLoS ONE* 5:e15554. doi: 10.1371/journal.pone.0015554

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to The Bootstrap.* London: Chapman & Hall. doi: 10.1007/978-1-4899-4541-9

Fletcher, P. C., and Frith, C. D. (2008). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* 10, 48–58. doi: 10.1038/nrn2536

Flor, H., Birbaumer, N., Hermann, C., Ziegler, S., and Patrick, C. J. (2002). Aversive pavlovian conditioning in psychopaths: peripheral and central correlates. *Psychophysiology* 39, 505–518. doi: 10.1111/1469-8986.3940505

Frank, M. J., Seeberger, L. C., and O'Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940–1943. doi: 10.1126/science.1102941

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *Neuroimage* 61, 1300–1310. doi: 10.1016/j.neuroimage.2012.04.018

Friston, K. (2013). Sample size and the fallacies of classical inference. *Neuroimage* 81, 503–504. doi: 10.1016/j.neuroimage.2013.02.057

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Pro. Natl. Acad. Sci. U.S.A.* 105, 6741–6746. doi: 10.1073/pnas.0711099105

Hare, R. D. (2003). *Manual for the Revised Psychopathy Checklist, 2nd Edn.* Toronto, ON: Multi-Health Systems.

Hare, R. D., Hart, S. D., and Harpur, T. J. (1991). Psychopathy and the DSM-IV criteria for antisocial personality disorder. *J. Abnorm. Psychol.* 100, 391–398. doi: 10.1037/0021-843X.100.3.391

Hartmann, A., Van Der Kooij, A. J., and Zeeck, A. (2009). Exploring nonlinear relations: models of clinical decision making by regression with optimal scaling. *Psychother. Res.* 19, 482–492. doi: 10.1080/10503300902905939

Huys, Q. J. M., Moutoussis, M., and Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Netw.* 24, 544–551. doi: 10.1016/j.neunet.2011.03.001

Jelicic, M., Merckelbach, H., Timmermans, M., and Candel, I. (2004). De nederlandstalige versie van de psychopathic personality inventory: enkele psychometrische bevindingen [The dutch version of the psychopathy personality inventory: some psychometric results]. *De Psycholoog* 12, 604–608.

Jones, A. P., Laurens, K. R., Herba, C. M., Barker, G. J., and Viding, E. (2009). Amygdala hypoactivity to fearful faces in boys with conduct problems and callous-unemotional traits. *Am. J. Psychiatry* 166, 95–102. doi: 10.1176/appi.ajp.2008.07071050

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., and Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science* 321, 806–810. doi: 10.1126/science.1156902

Lilienfeld, S. O., and Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in non-criminal populations. *J. Pers. Assess.* 66, 488–524. doi: 10.1207/s15327752jpa6603_3

Lilienfeld, S. O., Patrick, C. J., Benning, S. D., Berg, J., Sellbom, M., and Edens, J. F. (2012). The role of fearless dominance in psychopathy: confusions, controversies, and clarifications. *Personal. Disord.* 3, 327–340. doi: 10.1037/a0026987

Maia, T. V., and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* 14, 154–162. doi: 10.1038/nn.2723

Mars, R. B., Shea, N. J., Kolling, N., and Rushworth, M. F. (2012). Model-based analyses: promises, pitfalls, and example applications to the study of cognitive control. *Q. J. Exp. Psychol.* 65, 252–267. doi: 10.1080/17470211003668272

Miller, J. D., and Lynam, D. R. (2012). An examination of the Psychopathic Personality Inventory's nomological network: a meta-analytic review. *Personal. Disord.* 3, 305. doi: 10.1037/a0024567

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–78. doi: 10.1016/j.tics.2011.11.018

Neumann, C. S., Schmitt, D. S., Carter, R., Embley, I., and Hare, R. D. (2012). Psychopathic Traits in Females and Males across the Globe. *Behav. Sci. Law* 30, 557–574. doi: 10.1002/bsl.2038

Newman, J. P., and Kosson, D. S. (1986). Passive avoidance learning in psychopathic and nonpsychopathic offenders. *J. Abnorm. Psychol.* 95, 252–256. doi: 10.1037/0021-843X.95.3.252

Newman, J. P., Patterson, C. M., Howland, E. W., and Nichols, S. L. (1990). Passive avoidance in psychopaths: the effects of reward. *Pers. Individ. Dif.* 11, 1101–1114. doi: 10.1016/0191-8869(90)90021-I

Nieuwenhuis, S., Gilzenrat, M. S., Holmes, B. D., and Cohen, J. D. (2005). The role of the locus coeruleus in mediating the attentional blink: a neurocomputational theory. *J. Exp. Psychol.* 134, 291. doi: 10.1037/0096-3445.134.3.291

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454. doi: 10.1126/science.1094285

Patrick, C. J., Bradley, M. M., and Lang, P. J. (1993). Emotion in the criminal psychopath: startle reflex modulation. *J. Abnorm. Psychol.* 102, 82–92. doi: 10.1037/0021-843X.102.1.82

Patrick, C. J., Fowles, D. C., and Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: developmental origins of disinhibition, boldness, and meanness. *Dev. Psychopathol.* 21, 913–938. doi: 10.1017/S0954579409000492

Paulhus, D. L., Neumann, C. S., and Hare, R. D. (2013). *Manual for the Hare Self-Report Psychopathy Scale*. Toronto, ON: Multi-Health Systems.

Poythress, N. G., Edens, J. F., and Lilienfeld, S. O. (1998). Criterion-related validity of the psychopathic personality inventory in a prison sample. *Psychol. Assess.* 10, 426. doi: 10.1037/1040-3590.10.4.426

Redish, A. D., Jensen, S., and Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *Behav. Brain Sci.* 31, 415–436. doi: 10.1017/S0140525X0800472X

Sadeh, N., and Verona, E. (2008). Psychopathic personality traits associated with abnormal selective attention and impaired cognitive control. *Neuropsychology* 22, 669–680. doi: 10.1037/a0012692

Seara-Cardoso, A., Dolberg, H., Neumann, C., Roiser, J. P., and Viding, E. (2013). Empathy, morality and psychopathic traits in women. *Pers. Individ. Dif.*, 55, 328–333. doi: 10.1016/j.paid.2013.03.011

Seara-Cardoso, A., Neumann, C., Roiser, J., McCrory, E., and Viding, E. (2012). Investigating associations between empathy, morality and psychopathic personality traits in the general population. *Pers. Individ. Dif.* 52, 67–71. doi: 10.1016/j.paid.2011.08.029

Sellbom, M., Ben-Porath, Y. S., Lilienfeld, S. O., Patrick, C. J., and Graham, J. R. (2005). Assessing psychopathic personality traits with the MMPI–2. *J. Pers. Assess.* 85, 334–343. doi: 10.1207/s15327752jpa8503_10

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.1467-9868.2011.00771.x

Verona, E., Sprague, J., and Sadeh, N. (2012). Inhibitory control and negative emotional processing in psychopathy and antisocial personality disorder. *J. Abnorm. Psychol.* 121, 498–510. doi: 10.1037/a0025308

Viding, E., Frick, P. J., and Plomin, R. (2007). Aetiology of the relationship between callous-unemotional traits and conduct problems in childhood. *Br. J. Psychiatry* 190, s33–s38. doi: 10.1192/bjp.190.5.s33

von Borries, A. K. L., Brazil, I. A., Bulten, B. H., Buitelaar, J. K., Verkes, R. J., and de Bruijn, E. R. A. (2010). Neural correlates of error-related learning deficits in individuals with psychopathy. *Psychol. Med.* 40, 1443–1451. doi: 10.1017/S0033291709992017

White, S. F., Pope, K., Sinclair, S., Fowler, K. A., Brislin, S. J., Williams, W. C., 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, 315–323. doi: 10.1176/appi.ajp.2012.12060840

Yoshida, W., Dolan, R. J., and Friston, K. J. (2008). Game theory of mind. *PLoS ONE Comput. Biol.* 4:e1000254. doi: 10.1371/journal.pcbi.1000254

Yu, A. J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron* 46, 681–692. doi: 10.1016/j.neuron.2005.04.026