Transdiagnostic phenotyping reveals a range of metacognitive deficits associated with compulsivity

X.F. Tricia Seow[1,2] and Claire M. Gillan[1,2,3]

[1]*Department of Psychology, Trinity College Dublin, Dublin, Ireland*

[2]*Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland*

[3]*Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland*

Abstract

Recent work suggests that obsessive-compulsive disorder (OCD) is underpinned by a breakdown in the relationship between explicit beliefs about how the world works (i.e. confidence about states) and updates to behaviour. The present study aimed to test the precise computations that underlie this disconnection and ascertain their specificity to OCD symptoms. We phenotyped a large general population sample (N = 437), who also completed a predictive inference task and found that decreases in action-confidence coupling were not only associated with OCD symptoms, but also several other clinical phenotypes (6/9 corrected for multiple comparisons). This non-specific pattern was explained by a transdiagnostic compulsive symptom dimension. Action-confidence decoupling in high compulsives was associated with a marked inability to update confidence estimates (but not behaviour) according to unexpected outcomes, uncertainty and positive feedback. Our findings suggest that compulsive behaviour may be explained by difficulty in building an accurate explicit model of the world.

Introduction

Intentional decisions are dependent on the interplay between behaviour and beliefs. Beliefs guide behaviour, and the consequences of our behaviour in turn update beliefs. Computational models of learning suggest that the strength of belief (i.e. “confidence”) governs the extent of its influence on action; the less confident we are, the less our behaviour is influenced by pre-existing beliefs, compared to new information (Behrens, Woolrich, Walton, & Rushworth, 2007; Nassar, Wilson, Heasly, & Gold, 2010). It has been suggested that a breakdown in the relationship between action and belief is characteristic of compulsive behaviours, for example in obsessive-compulsive disorder (OCD) or addiction. In these disorders, behaviour often appear autonomous, unguided by conscious control or ‘ego-dystonic’ – for instance, persistent drug use despite negative consequences (Everitt & Robbins, 2005) or out-of-control repetitive checking despite knowing the door has been locked (Fineberg et al., 2010). One potential cause of the divergence between intention and action in compulsive individuals is an impairment in the brain’s goal-directed system, which links actions to consequences and protects against overreliance on rigid habits (Gillan, Otto, Phelps, & Daw, 2015). Deficits in goal-directed planning have been consistently observed in OCD (Gillan, Apergis-Schoute, et al., 2015; Gillan, Morein-Zamir, Kaser, et al., 2014; Gillan, Morein-Zamir, Urcelay, et al., 2014; Gillan et al., 2011) and related disorders (Voon et al., 2015) and there is evidence to suggest this constitutes a transdiagnostic psychiatric trait linked to several aspects of clinically-relevant compulsive behaviour (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016).

Despite consistency of findings, the precise mechanism supporting this dysfunction is only partially understood. This is because most tasks that have been employed struggle to separate the construction of an internal model (e.g. action-outcome knowledge) from its implementation in behaviour. Those that have attempted this have yielded interesting, if equivocal, results. One study showed that OCD patients get stuck in habits, even when they possess the requisite action-outcome knowledge to theoretically perform in a goal-directed fashion (Gillan, Morein-Zamir, Urcelay, et al., 2014). This suggests that the implementation of goal-directed behaviour is deficient in OCD, independent of their ability to construct of the model. But this does not mean the internal model is intact; studies that used more challenging paradigms have found

deficits in the acquisition of explicit action-outcome contingency knowledge itself in OCD patients (Gillan et al., 2016), suggesting that patients may have problems with both. These findings come from paradigms where instrumental action typically affects the kind of information that is gathered and as such are somewhat confounded and difficult to interpret. Recently Vaghi and colleagues addressed this metacognitive question in OCD patients with more precision – using a paradigm that examined how patients make trial-wise adjustments to behaviour (i.e. implicit model) and confidence (i.e. explicit model) in response to feedback (Nassar et al., 2016; Vaghi et al., 2017). They found that in OCD, the association between confidence and behavioural updating ('action-confidence coupling') was diminished – in other words, patients' behaviour did not align with their internal model. They probed this further and found that while confidence estimates did not differ from healthy controls, OCD patients showed abnormalities in their behavioural learning rate, making more trial-wise adjustments in response to feedback than healthy controls (Vaghi et al., 2017).

The finding of intact confidence in OCD is consistent with a prior study where 40 individuals high versus low in OCD symptoms had no differences in their confidence in their perceptual decision-making performance (Hauser, Allen, Consortium, Rees, & Dolan, 2017). It also aligns with findings from two large internet-based samples (each N > 490) using the same task, which showed that OCD symptoms had no association with confidence (Rouault, Seow, Gillan, & Fleming, 2018). One problem with this kind of study design and analysis, however, is that it fails to capture the potentially competing influence of co-occurring disorders/symptoms in psychiatric populations. Even in studies where certain comorbid diagnoses are explicitly excluded for, as in Vaghi *et al.*, rates of depression and anxiety are in excess of those of controls (Vaghi et al., 2017). Similarly, when severity of self-report anxiety and depression are matched across groups by design, as in Hauser *et al.* (Hauser et al., 2017), this does not accurately reflect the average OCD patient where comorbidity is the rule, not the exception, in psychiatry, with for example >25% of patients with OCD meeting criteria for 4 or more additional diagnoses (Gillan, Fineberg, & Robbins, 2017). Indeed, selecting for individuals with high OCD scores but low depression scores from a large sample might reflect statistical anomalies rather than a true and lasting absence of depression in those individuals. An alternative approach is one that measures these relevant co-occurring symptoms in a large sample and seeks to account for their

(competing or inflating) influence on the cognitive measure of interest. We took this approach in a prior study and found that abnormalities in confidence (which were not reliably linked to either OCD or depressive symptoms in this same sample) were robustly associated with two transdiagnostic psychiatric dimensions – in opposing directions. 'Anxious-depression' was associated with reduced confidence, while 'compulsive behaviour and intrusive thought' was linked to inflated confidence.

This methodology seeks to address the existential issue with diagnoses in psychiatry (Huys, Moutoussis, & Williams, 2011; Hyman, 2007; Insel et al., 2010) and develop new ways of capturing (and defining) meaningful transdiagnostic psychiatric dimensions that bear a closer relation to biologically or cognitively important processes (Friston, Stephan, Montague, & Dolan, 2014; Gillan & Daw, 2016; Huys, Maia, & Frank, 2016; Stephan & Mathys, 2014). In the studies published to date, this method has demonstrated that transdiagnostic traits map cognition with greater specificity and strength of association than questionnaires based on DSM disorders in several domains (Gillan et al., 2016; Patzelt, Kool, Millner, & Gershman, 2019; Rouault et al., 2018).

In the present study, we used this method to test the relationship between transdiagnostic symptom dimensions and action-confidence coupling. Using the same task from Vaghi and colleagues (Vaghi et al., 2017), we related self-report psychiatric symptoms to the extent to which trial-wise adjustments in action related to confidence reports in a sample of 437 subjects recruited from Amazon's Mechanical Turk. Consistent with Vaghi and colleagues, we found that OCD symptoms correlated with a decrease in action-confidence coupling. However, this association was by no means specific to OCD – deficits in action-confidence coupling were associated with scores on 6/9 of the different questionnaires included in this study. This was after imposing a strict correction for multiple comparisons (it was correlated to all 9, if no correction was applied). Using transdiagnostic dimensions, a considerably more specific result was evident; only scores on the 'compulsive behaviour and intrusive thought' dimension was related to the decoupling of confidence and behaviour.

Next, we tested if this decoupling of action and confidence arose from failures in action-updating (Vaghi et al., 2017) or confidence (Rouault et al., 2018). We found

evidence for the latter; there was no difference in action updating associated with either OCD symptoms or the compulsive dimension, but both OCD symptoms and compulsivity were linked to increases in confidence. As expected, we observed that confidence showed a double dissociation with anxious-depression and the compulsive dimension, where anxious-depression was associated with reduced confidence (Rouault et al., 2018). These findings replicate and extend prior work (Rouault et al., 2018) that used a perceptual decision-making task by showing that these effects are also evident when confidence pertains to reinforcement learning. They also illustrate how findings from traditional OCD versus healthy control group-level comparisons might obscure specific and important associations due to the competing influence of comorbid symptoms, in this case, anxious-depression.

Lastly, we investigated if individuals high in compulsivity had abnormalities in the way they used information to update confidence, with a view to uncovering the source of inflated confidence. With a reduced Bayesian model (McGuire, Nassar, Gold, & Kable, 2014; Nassar et al., 2016; Nassar et al., 2010; Vaghi et al., 2017), we found that confidence was not just elevated in high compulsive individuals, but was also less responsive to various parameters that should drive learning such as unexpected outcomes, environmental uncertainty and positive feedback. Importantly, compulsive individuals had no difficulty updating behaviour in response to the same sources of evidence. The lack of an effect of evidence on action updating is consistent with prior work showing that feedback-based instrumental learning is not impaired in OCD (Gillan et al., 2016; Voon et al., 2015), but in contrast to the findings of Vaghi and colleagues. Further investigation suggested that this difference was not explained by an advantage of the transdiagnostic approach *per se*, but likely due to our ability to control for important demographic characteristics in a larger sample.

In summary, compulsivity is linked to problems in developing an explicit and accurate model of the decision space, and this might contribute to broader class of problems these individuals face with goal-directed planning and execution. A transdiagnostic approach resolves the apparent generalisability of action-confidence coupling deficits in psychiatry and reveals an exclusive association with compulsivity.

Results

Participants (N = 437) performed a predictive-inference task in which their goal was to catch a flying particle by placing a bucket in the correct location (Figure 1). On each trial, they repositioned the bucket and rated how confident they were in catching the particle. After the behavioural task, participants completed an IQ test and a battery of self-report questionnaires assessing a range of psychiatric symptoms (see Methods). Individual item-level responses on these questionnaires were transformed into three transdiagnostic dimensions using loadings defined in prior study (Gillan et al., 2016), 'Anxious-Depression' (AD), 'Compulsive Behaviours and Intrusive Thought' (CIT) and 'Social Withdrawal' (SW).

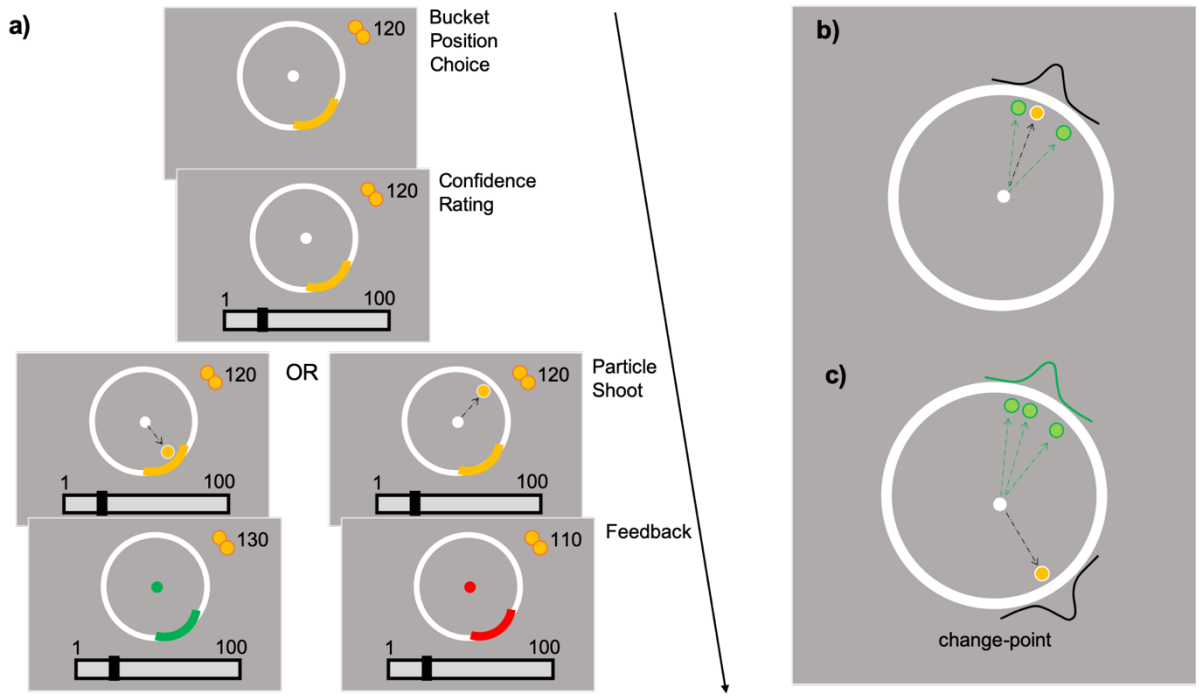

Figure 1. *Predictive-inference task.*
(a) Trial sequence. *Participants were instructed to position a bucket (yellow arc on the circle edge) to catch a flying particle, and thereafter rated their confidence that they would catch the particle. Particles were fired from the center of the circle to the edge. Points were gained when the particle was caught, and the bucket turned green; else, points were lost and the bucket turned red.*
(b) Particle trajectories. *For every trial, landing locations were independently sampled from a Gaussian distribution. As such, particles landed around the same area with small variations induced by noise. For illustration purposes, dashed arrow lines represent particle trajectory of current (black) and past (green) trials, which over trials allow subjects to generate a representation of the Gaussian.*
(c) Change-points. *The mean of the distribution abruptly moves to another point on the circle when a "change-point" occurs. This new mean is then sampled in the same manner as* ***(b)*** *until the next change-point.*

Action-confidence decoupling is linked to various psychiatric symptoms. In line with prior research, size of action updates (bucket position difference from trial *t* and *t*+1) were strongly related to confidence within-subjects (β = -8.85, *Standard Error (SE)* = 0.11, $p < 0.001$, Figure 2a), such that lower confidence was linked to larger updates, in the sample as a whole. Previous work by Vaghi *et al.* found that OCD patients exhibited less coupling between action and confidence compared to controls, which was correlated to the severity of self-reported OCD symptomology within the patient sample (Vaghi et al., 2017). We tested the latter in a general population sample and replicated this result; the severity of OCD symptoms was associated with significantly lower action-confidence coupling (β = 1.30, *SE* = 0.21, $p < 0.001$, corrected., Figure 2b). However, we found that this relationship was not specific – all nine questionnaires of varied psychiatric symptom severity showed a similar pattern of reduced coupling (6/9 questionnaires (alcohol addiction, depression, eating disorders, impulsivity, OCD and schizotypy) had significant decoupling at $p < 0.001$ corrected; the remaining three (apathy, social anxiety, trait anxiety) trended in the same direction, but did not survive Bonferroni correction for multiple comparisons).

Transdiagnostic analysis shows a more specific pattern. When we refactored the data into three transdiagnostic dimensions defined previously in the literature, a profoundly different picture emerged. CIT was the only dimension to show decreased action-confidence coupling (β = 1.57, *SE* = 0.23, $p < 0.001$, corrected). Together this suggests that reductions in action-confidence coupling are not unique to OCD, but rather better represent a feature of compulsivity more broadly.

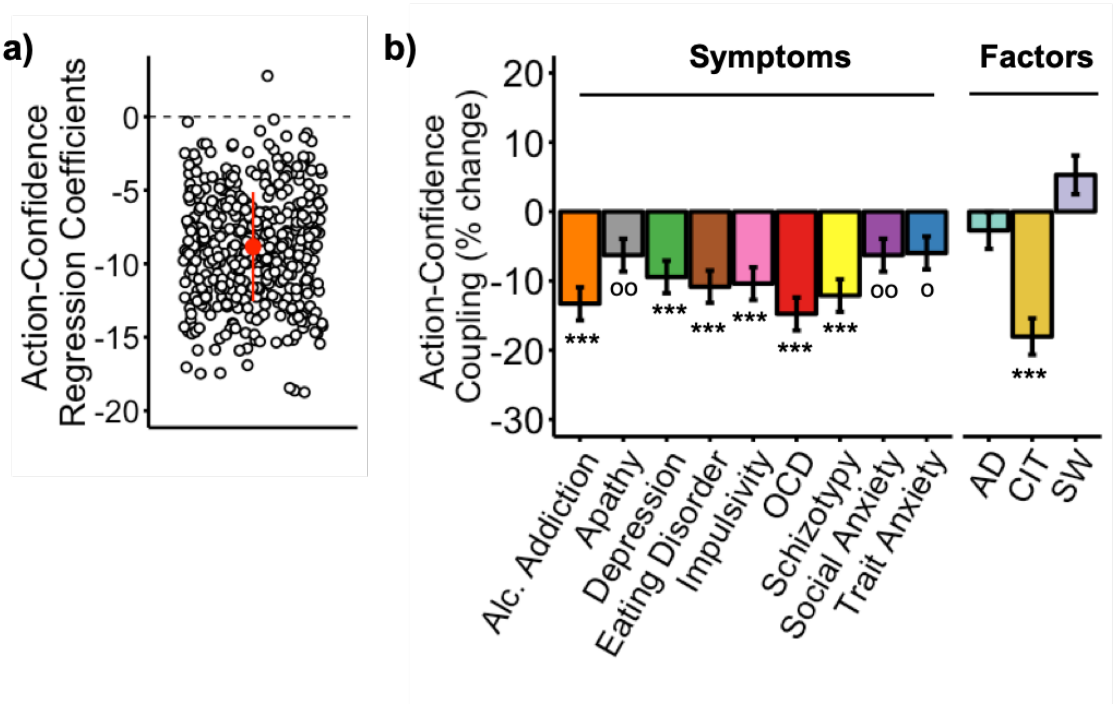

Figure 2. *Action-confidence coupling and its relationship with psychiatric symptoms and factors (controlled for IQ, age and gender).*
(a) *Regression model where action update is predicted by confidence. Individual coefficients are represented by white dots. Red marker indicates the mean and standard deviation. As expected, regression coefficients were negative, such that higher confidence was associated with smaller updates to the bucket position ('action').*
(b) *Associations between action-confidence coupling and self-reported psychopathology or psychiatric factors. All symptoms predicted a decrease in action-confidence coupling. This decoupling relationship was specifically captured by the 'Compulsive Behaviour and Intrusive Thought' dimension with its effect size larger than for any individual questionnaire. Each psychiatric symptom was examined in a separate regression, while factors were included in the same regression model. The Y-axes shows the percentage decrease in the size of the action-confidence coupling effect as a function of 1 standard deviation increase of symptom/factor scores. Error bars denote standard errors.* $^{o}p < 0.05$, $^{oo}p < 0.01$ *uncorrected,* $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. *Results are Bonferroni corrected for multiple comparisons over number of symptoms/factors.*

Compulsivity is linked to inflated confidence, not aberrant action updating. CIT was associated with higher overall confidence levels (β = 6.74, SE = 1.02, $p < 0.001$, corrected), while AD was associated with lower confidence (β = -3.42, SE = 1.04, p = 0.003, corrected) (Figure 3a). This constitutes a full replication of prior work, which assessed confidence in the context of low-level perceptual decision-making(Rouault et al., 2018). Our findings thus illustrate that these bidirectional associations with confidence are not only robust, but generalisable to confidence applied to other aspects of the decision-system – i.e. reinforcement learning. In contrast to confidence, CIT was not associated with changes in action updating, i.e. the overall tendency to move the bucket from trial to trial. Only SW showed an association with action

updating, such that participants scoring high in this dimension tended to move the bucket more (β = 0.89, *SE* = 0.28, *p* = 0.005, corrected, Figure 3b).

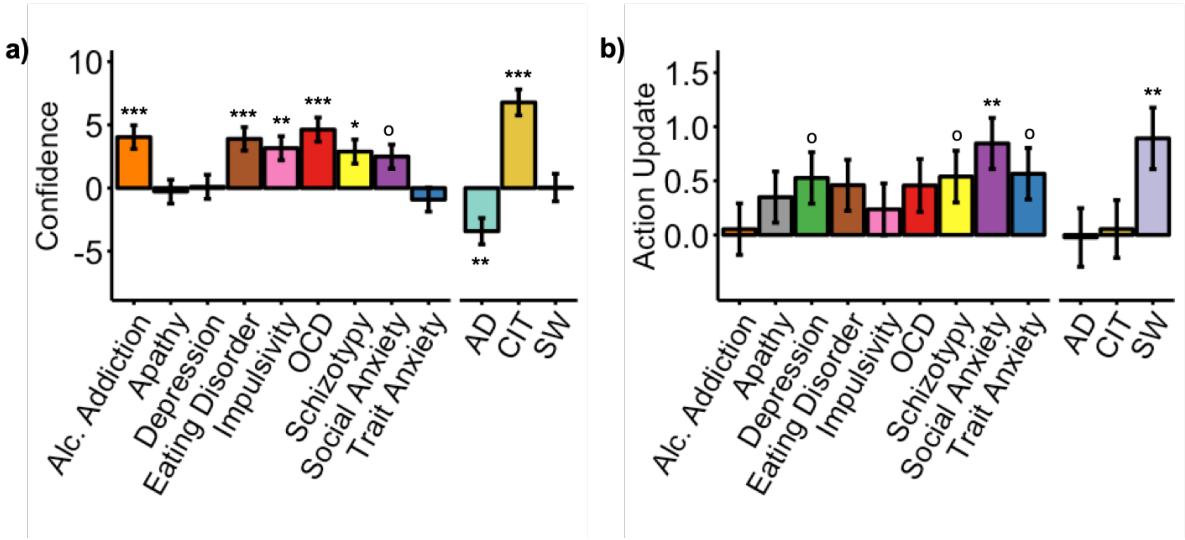

Figure 3. *Associations between psychiatric symptoms, or transdiagnostic factors (controlled for IQ, age and gender) with:*
(a) *Confidence rating on each trial. Most of the questionnaires scores were positively associated with confidence. However, refactoring into transdiagnostic traits revealed previously obscured bidirectional associations. 'Anxious-depression' (AD) was linked to decreased confidence, while 'Compulsive Behaviour and Intrusive Thought' (CIT) was linked to increased confidence.*
(b) *Action updates (i.e. bucket movement from one trial to the next). Only social anxiety was associated with an increased tendency to move the bucket, and this was similarly captured by, and specific to, the 'Social Withdrawal' (SW) factor.*
The Y-axes shows the percentage decrease in the size of the action-confidence coupling effect as a function of 1 standard deviation increase of symptom/factor scores. Error bars denote standard errors. $^{o}p < 0.05$, $^{oo}p < 0.01$ uncorrected, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Results are Bonferroni corrected for multiple comparisons over number of symptoms/factors.

Confidence in compulsivity is less sensitive to unexpected outcomes, environment uncertainty and positive feedback. The previous analyses suggest that confidence in compulsive individuals is both inflated and decoupled to behaviour. To understand the mechanism behind this, we tested the extent to which confidence estimates were sensitive to multiple factors that should drive belief-updating. Specifically, prior work has shown that trial-wise adjustments in behaviour are be influenced by the 1) information gained from the most recent change in particle location, 2) surprising large particle location changes owing to change-points and the 3) uncertainty of one's belief about the particle landing location distribution mean (McGuire et al., 2014). To separate the contributions of these factors, we computed three normative parameters with a quasi-optimal Bayesian model (previously

employed in Vaghi *et al.* (Vaghi et al., 2017) and validated in prior studies (McGuire et al., 2014; Nassar et al., 2016; Nassar et al., 2010), with further model details in the online supplement) to the sequence of particle locations experienced by each participant. The parameters of the model included PE^b (model prediction error, the tendency to update towards the most recent particle landing location), CPP (change-point probability, the likelihood that a surprising outcome had occurred) and RU (relative uncertainty owing to the imprecise estimation of the distribution mean based on previous outcomes).

We analysed trial-wise confidence using regression models with these parameters including a categorial Hit regressor (previous trial was a hit or miss), and controlled for age, gender and IQ. In line with Vaghi *et al.* (Vaghi et al., 2017), we found that confidence was influenced by PE, CPP, RU and Hit (Table S1). The CIT symptom dimension was associated with a significantly diminished influence of CPP (β = 0.05, SE = 0.01, p < 0.001, corrected), RU (β = -0.05, SE = 0.01, p < 0.001, corrected) and Hit (β = -0.03, SE = 0.01, p = 0.003, corrected) on confidence (Figure 4a & Figure S5). In other words, confidence estimates in CIT were less sensitive to unexpected outcomes, the uncertainty of the true distribution mean and whether the bucket position on the previous trial was correct (i.e. the particle was caught). These results suggest that confidence in highly compulsive individuals is not only inflated, it is also only weakly connected to several sources of environmental evidence. There were no associations between AD or SW and the extent to which this evidence influenced confidence (Figure S5).

Action updates in compulsivity respond appropriately to evidence. Using the same approach described for confidence, trial-wise adjustments in action (the absolute difference from bucket position on trial t and t+1) were influenced by all model parameters (Table S1). In contrast to confidence, CIT was not linked to changes in the influence of any of the parameters on action (Figure S5). SW was related to a significant increased influence of PE^b, suggesting that individuals high in this trait had an increased tendency to update their action with every new outcome (β = -0.06, SE = 0.02, p < 0.05, corrected) (Figure S5). There were no associations with AD. Additional analyses in the online supplement show that when demographics are not

controlled for, some apparent associations between action-updating and compulsivity emerge that correspond to those reported previously in OCD (Vaghi et al., 2017).

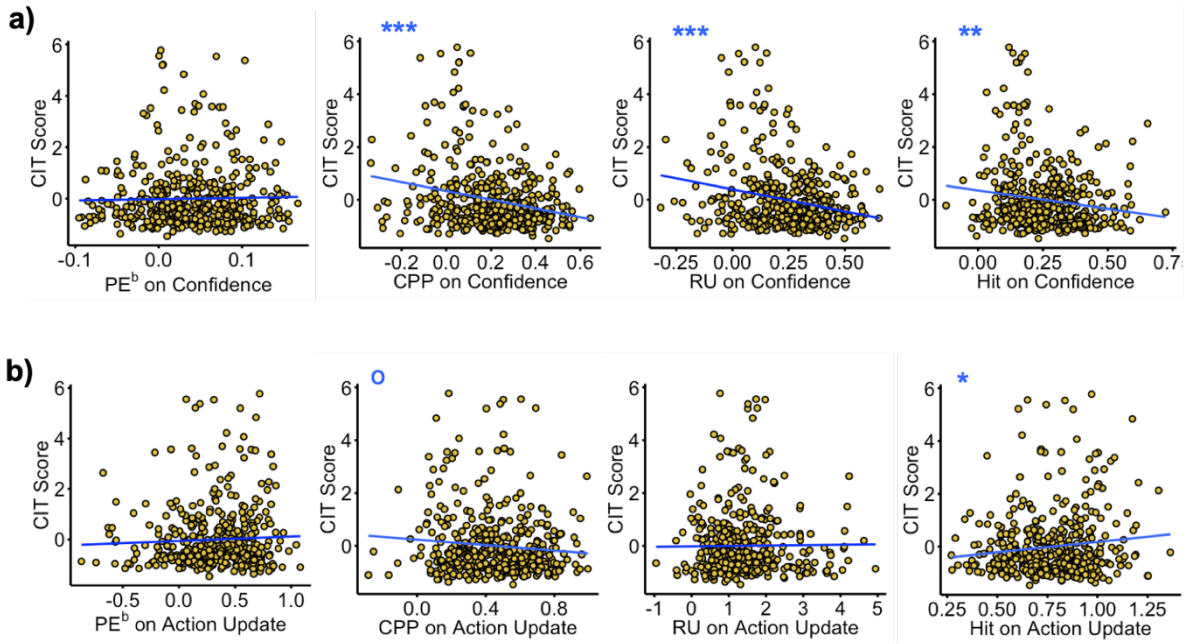

Figure 4. *Confidence level/action update was predicted by absolute model prediction error (PE^b), change-point probability (CPP), relative uncertainty (RU) and hit/miss categorial regressor (Hit), controlled for IQ, age and gender. Coefficient estimates from the model were correlated with 'Compulsive Behaviour and Intrusive Thought' (CIT) severity.*
(a) *CIT was found to be associated with significantly diminished influence of CPP, RU and Hit on confidence. PE^b, CPP and RU on confidence coefficients are inverted to illustrate direction of effects. PE^b: $r = 0.02$, $p = 1.00$; CPP: $r = -0.24$, $p < 0.001$; RU: $r = -0.24$, $p < 0.001$; Hit: $r = -0.15$, $p = 0.005$.*
(b) *In contrast, CIT was found not linked to changes in the influence of any of model parameters on action update. For plotting purposes, we show the association of parameter and compulsivity without controlling for AD and SW. PE^b: $r = 0.05$, $p = 0.98$; CPP: $r = -0.10$, $p = 0.13$; RU: $r = 0.01$, $p = 1.00$; Hit: $r = 0.12$, $p = 0.03$. Note that a modest association with Hit on action update is observed here and illustrated in the associated plot, but does not survive inclusion of all three factors in the same model.*
*Dots represent coefficients of individual participants for model parameters from a basic model of confidence/action update ~ regressors*demographics (x-axis), against their CIT score (y-axis) (see Methods). Hit on action update coefficients are inverted to illustrate direction of effects, such that CIT is linked to an increase influence of hits on action updating (which is negative in direction). $^{o}p < 0.05$, uncorrected, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Results are Bonferroni corrected for multiple comparisons over the three factors.*

Discussion

In this study, we demonstrated that a breakdown in the relationship between explicit belief (confidence) and behaviour is associated with a transdiagnostic psychiatric dimension - compulsive behaviour and intrusive thought (CIT). This decoupling arises from abnormalities in belief, rather than behaviour. Individuals high in CIT exhibited inflated confidence estimates on average, as well as failures in utilising unexpected outcomes, belief uncertainty and positive feedback to inform their confidence levels appropriately. In contrast, action updating in response to these factors did not differ as a function of severity of CIT. Our findings implicate a dysfunctional metacognitive mechanism in compulsivity, whereby those high on the dimension have difficulty in updating their explicit model of the world in response to various sources of evidence.

Existing models of compulsivity propose that deficits in goal-directed control leave individuals vulnerable to establishing compulsive habits (Gillan & Robbins, 2014). Evidence for this has primarily come from behavioural tests, in which patients with OCD (and other compulsive disorders) have difficulty in exerting control over well-trained habits when motivations change (i.e. a devaluation test) (Ersche et al., 2016; Gillan, Apergis-Schoute, et al., 2015; Gillan, Morein-Zamir, Urcelay, et al., 2014; Gillan et al., 2011). Other tasks have shown that compulsive patients have deficits in using a model of the world to make choices in a prospective fashion (even when habits are not present), relying instead solely on reinforcement (i.e. feedback) to direct choice (Gillan, Morein-Zamir, Kaser, et al., 2014; Voon et al., 2014). The finding of the present study, that individuals high in compulsivity fail to update their world-model in response to several types of evidence, is an important extension of this literature. It has until now been presumed that the challenge facing compulsive individuals was in the implementation of the model, rather than its generation and/or maintenance (Gillan & Robbins, 2014). This has implications not just for our understanding of compulsive disorders, but also their treatment. Recent work has suggested that metacognitive skills can be improved though adaptive training (Carpenter et al., 2019), suggesting there is potential for this kind of treatment in psychiatric populations where compulsivity is an issue.

Confidence was not just unresponsive to various factors underlying learning, it was also inflated in compulsive individuals. This finding replicates prior work that examined confidence in the context of perceptual decision making (Rouault et al., 2018), showing it extends to reinforcement learning and thus appears to be domain-general. Future work will be needed to dissect the specific mechanism through which confidence becomes inflated in compulsivity. In instances when confidence diverges from action, prior work has suggested confidence estimates may be corrupted by noise, internal states or a continued/lack of evidence processing (Fleming & Daw, 2017; Meyniel, Sigman, & Mainen, 2015). Coupled with the finding that confidence is less informed by several sources of evidence in individuals high in compulsivity, it is possible that inflated confidence in compulsivity arises through some unmodeled source of information or noise. This posit is supported by the finding that actions were updated normally in response to feedback in high compulsives, which accords with prior work studying basic reinforcement learning in compulsive patient groups (i.e. 'model-free' learning) that found it to be intact (Gillan et al., 2016; Voon et al., 2015). That said, our findings are in conflict with a previous study that found increased action-updating tendencies in OCD (Vaghi et al., 2017). In this case, the discrepancy does not appear to be explained by the putative superiority of a transdiagnostic approach, but rather additional analyses in the online supplement suggest this difference might be explained by demographic confounds, which our study was sufficiently powered to control for.

Beyond the specific results of this study with respect to confidence and compulsivity, these data highlight the benefit of transdiagnostic dimensions over traditional modes of phenotyping. When we examined questionnaires that are ubiquitous in clinical research, but rarely compared to one another, we found strikingly non-specific patterns of association with task variables. For example, all nine questionnaires showed an association with action-confidence coupling in the same direction (6/9 surviving strict correction). In contrast, the compulsive factor was the only transdiagnostic dimension to show an association. In addition to resolving issues with collinearity across questionnaires, this approach also resolves issues associated with the heterogeneity within them. For example, severity of neither depression nor anxiety was associated with decreases in confidence using a standard clinical questionnaire, but the anxious-depression dimension was. In comparison to work with diagnosed

patients, the benefits of the transdiagnostic approach are the same. Vaghi and colleagues found no difference in OCD patients' mean confidence ratings compared to healthy controls (Vaghi et al., 2017), while we found a strong a reproducible association between compulsivity and inflated confidence and anxious-depression and diminished confidence (Rouault et al., 2018). Given that OCD is frequently co-morbid with anxiety disorders (over 75% (Ruscio, Stein, Chiu, & Kessler, 2010)), which has an opposing relationship to confidence, it is no surprise that differences between OCD patients and controls are not detectible. Together, these data suggest that transdiagnostic phenotyping may provide a closer fit to underlying brain processes than DSM distinctions.

This study was not without limitation. All of our measures were taken online via Amazon's Mechanical Turk and so experimenter control of the testing environment was virtually non-existent. Additionally, as the task was adapted for web-based testing, bucket movements were controlled by keyboard presses as opposed to a rotor controller as in Vaghi *et al.* (Vaghi et al., 2017), which may contribute to increased noise in spatial update measure. However, our ability to collect a large sample was evidently sufficient to mitigate the issue of increased noise in data collected online. For example, we were able to reproduce basic main effects similar to Vaghi *et al.* (Vaghi et al., 2017) (Table S1) and we also replicated previously observed associations with confidence, CIT and AD (Rouault et al., 2018). Even though we reproduced the factor structure from a prior paper in our own data (Figure S6), we nonetheless used the factor weights from this prior publication (Gillan et al., 2016) to transform raw questionnaire scores into transdiagnostic factors for analysis. This ensured independence and underscores the robustness and reproducibility of these factors and their association to cognition. The extent to which these results are applicable to diagnosed patients, as opposed to a general population sample, is not something we can directly address here and is thus another limitation. However, it is notable that we replicated the association between OCD symptoms and action-confidence decoupling observed in a clinical sample that were tested in-person (Vaghi et al., 2017). Similar results have been found for goal-directed planning, which is both deficient in patients tested in-person (Gillan et al., 2011), and correlated with OCD symptoms in the general population, tested online (Gillan et al., 2016; Snorrason, Lee, de Wit, & Woods, 2016). As such, there is no reason to suspect these findings are not

applicable to patients and as outlined above, this method has several strengths over the case-control approach. It directly addresses the issue of psychiatric co-morbidity, it helps us to achieve higher statistical power and thus promotes reproducibility, and finally, makes research faster, more efficient and more representative (Gillan & Daw, 2016).

To conclude, we highlighted how a transdiagnostic methodology can be crucial for uncovering specific associations between pathophysiology and clinical symptoms. We used this method to show that compulsive behaviour and intrusive thought is associated with reduced action-confidence coupling, inflated confidence and diminished influence of evidence on confidence estimates.

Materials and Methods

Power estimation. Previous research utilizing the predictive-inference task were constrained to small sample psychiatric populations (Vaghi et al., 2017). As such, we referred to earlier work that investigated confidence abnormalities in large general population cohorts with transdiagnostic symptom dimensions to determine an appropriate sample size (Rouault et al., 2018). The prior study reported an association of the 'Anxious-Depression' factor with lowered confidence level ($\beta = -0.20$, $p < 0.001$), an effect size suggesting that N = 295 participants were required to achieve 90% power at 0.001 significance level.

Participants. Data were collected online using Amazon's Mechanical Turk (AMT). Participants were 18 years or older, based in USA and had more than 95% of their previous tasks on AMT approved. After reading the study information and consent pages, they provided consent by clicking the 'I give my consent' button. Participants were paid a base sum of 7 USD. They could also earn a bonus of up to 1 USD depending on the number of points earned in the behavioural task. We collected data from N = 590 participants, of whom 249 were female (42.2%) with ages ranging from 20 to 65 (mean = 36.3, standard deviation (SD) = 10.2) years. All study procedures were approved by Trinity College Dublin School of Psychology Research Ethics Committee.

Exclusion criteria. Several exclusion criteria were applied to ensure data quality. Participants were excluded if they failed any of the following: (i) In the behavioural task, the confidence scale indicator would always start at either 25 or 75 on every trial. Participants who left their confidence rating as the default score for more than 60% of the trials (n > 180 trials) were excluded (N = 42). (ii) The task was also reset from the beginning if confidence ratings were left as the default score for >70% of the first 50 trials. 56 participants (9.82%) restarted the task at least once. Those who had their task reset >5 times were excluded (N = 6). (iii) Participants who had more than 50% correlation between the default score and their selected confidence rating were excluded (N = 109). (iv) Participants with a lower mean confidence where the previous trial was correct than incorrect were excluded (N = 66). (v) Participants who incorrectly responded to a "catch" question within the questionnaires: "If you are paying attention

to these questions, please select ‘A little’ as your answer” were excluded (N = 16). Combining all exclusion criteria, 153 participants (25.9%) were excluded. N = 437 participants were left for analysis.

Predictive inference task. We adapted the predictive-inference task from Vaghi *et al.* (Vaghi et al., 2017) for web-based testing (Figure 1). Left and right arrow keys enabled response navigation while a spacebar press was used for decision confirmation. The task consisted of a particle flying from the center of a large circle to its edge. First, participants positioned a ‘bucket’ (a free-moving arc on the circle edge) in order to catch the particle. Once bucket location was chosen, a confidence bar scaling 1 to 100 would appear below the circle after 500 ms. The confidence indicator would begin randomly at either 25 or 75. Participants then indicated how confident they were the particle would land in the bucket. After confirmation of the confidence report, a particle was then released from the center to fly towards the edge of the circle 800 ms later. If the particle landed within the boundaries of the bucket, the bucket would turn green for 500 ms and the participant gained 10 points; else, the bucket turned red for 500 ms and lost 10 points. The number of points accumulated over the task was presented in the top right-hand corner for participants to track their performance. Payment was performance contingent; the more points earned, the higher amount of bonus they received at the end, up to a maximum of 1 USD. Confidence ratings were not incentivized.

On each trial, the particle’s landing location on the circle edge was sampled independently and identically from a Gaussian distribution with SD = 12. As such, the particle landed in the same location with small variations determined by noise. The mean of this distribution did not change until a change-point trial was reached, where it was re-sampled from a uniform distribution U(1,360) (i.e. the number of points on the circle). Participants would therefore have to learn the mean of the new generative distribution after a change-point. The probability of a change-point occurring on each trial was determined by the hazard rate. In the task, there were two hazard rate conditions that varied the number of change-points in a stretch of 150 trials each: stable (hazard rate = 0.025, 4 change-points), and volatile (hazard rate = 0.125, 19 change-points). Hazard rate conditions were not relevant to the analyses of the current paper. The order of hazard rate conditions was randomly shuffled, as were the order

513 of change-points within a condition. Participants completed 300 trials in total, divided
514 into 4 blocks of 75 trials, with no explicit indication when a change in condition block
515 occurred. Breaks were given between blocks which did not fall before the switch of a
516 new hazard rate condition.
517
518 Before the start of the task, participants were instructed on the aim of the experiment
519 and shown its layout. Participants then completed 10 practice trials that were excluded
520 from the analysis and did not count for their final score. After the practice, they had to
521 answer 5 questions pertaining to the task. If they answered any of the questions
522 wrong, they would be brought back to the beginning of the instructions and taken
523 through the practice block again. Additionally, in order to reduce the number of
524 participants failing to utilize the confidence scale properly, the task was reset to the
525 beginning if participants left their confidence ratings as the default score for more than
526 70% of the trials at the 20th and 50th trial mark. They would have to answer the task
527 questions again before proceeding with the task.
528
529 **Self-report psychiatric questionnaires & IQ.** Participants completed a range of self-
530 report psychiatric assessments after the behavioural task. To enable application of the
531 transdiagnostic analysis with psychiatric dimensions described in previous studies
532 (Gillan et al., 2016; Rouault et al., 2018), we administered the same nine
533 questionnaires assessing: *Alcohol addiction* using the Alcohol Use Disorder
534 Identification Test (AUDIT) (Saunders, Aasland, Babor, Delafuente, & Grant, 1993),
535 *Apathy* using the Apathy Evaluation Scale (AES) (Marin, 1991), *Depression* using the
536 Self-Rating Depression Scale (SDS) (Zung, 1965), *Eating disorders* using the Eating
537 Attitudes Test (EAT-26) (Garner, Olmsted, Bohr, & Garfinkel, 1982), *Impulsivity* using
538 the Barratt Impulsivity Scale (BIS-10) (Patton, Stanford, & Barratt, 1995), *Obsessive-*
539 *compulsive disorder (OCD)* using the Obsessive-Compulsive Inventory – Revised
540 (OCI-R) (Foa et al., 2002), *Trait anxiety* using the trait portion of the State-Trait Anxiety
541 Inventory (STAI) (Spielberger, 1983), *Schizotypy* scores using the Short Scales for
542 Measuring Schizotypy (Mason, Linney, & Claridge, 2005), and *Social anxiety* using
543 the Liebowitz Social Anxiety Scale (LSAS) (Liebowitz, 1987). The order of these self-
544 report assessments administered was fully randomized. Following the questionnaires,
545 participants completed a Computerized Adaptive Task (CAT) based on items similar

546 to that of Raven's Standard Progressive Matrices (SPM) (Raven, 2000) to
547 approximate Intelligence Quotient (IQ).
548
549 **Transdiagnostic factors.** Raw scores on the 209 individual questions that subjects
550 answered from the 9 questionnaires were transformed into factor scores ('Anxious-
551 Depression', 'Compulsive Behaviour and Intrusive Thought', and 'Social Withdrawal'),
552 based on weights derived from a larger previous study (N = 1413) (Gillan et al., 2016).
553
554 **Action-confidence coupling.** First, we measured the coupling between action
555 updates (i.e. the tendency to move the bucket) and confidence. *Action* (the absolute
556 difference of bucket position on trial *t* and *t*+1) was the dependent variable and
557 *Confidence* (confidence level on trial *t*+1) was the independent variable in a trial-by-
558 trial regression analysis with age, gender and IQ as fixed effects co-variates (as with
559 all subsequent analyses). Within-subject factors (the intercept and main effect of
560 *Confidence*) were taken as random effects (i.e., allowed to vary across subjects).
561 *Confidence* was z-scored within-participant, while the fixed effect predictors were z-
562 scored across participant. If action and confidence are appropriately coupled,
563 participants should move the bucket more (larger *Action*) when their confidence levels
564 were low, producing a significant negative main effect of *Confidence* on *Action*. In the
565 syntax of the *lmer* function, the regression was: Action ~ Confidence * (Age + IQ +
566 Gender) + (1 + Confidence | Subject). Regression analyses were conducted using
567 mixed-effects models written in R, version 3.5.1 via RStudio
568 version 1.1.463 (http://cran.us.r-project.org) with the *lme4* package.
569
570 We then tested if psychiatric symptom severity was associated to changes in action-
571 confidence coupling by including the total score for each questionnaire (*Symptom*, z-
572 scored) as a between-subjects predictor in the model above. Separate regressions
573 were performed for each individual symptom due to high correlations across the
574 different psychiatric questionnaires. The extent to which questionnaire total scores
575 contribute to changes in action-confidence coupling is indicated by the presence of a
576 significant *Confidence***QuestionnaireScore* interaction. A positive interaction effect
577 indicates decreased action-confidence coupling (i.e., decoupling), while a negative
578 interaction effect indicates greater action-confidence coupling. The model was
579 specified as: Action ~ Confidence * (*QuestionnaireScore* + Age + IQ + Gender) + (1 +

Confidence | Subject). For the transdiagnostic analysis, we included all three factors in the same model, as correlation across variables was lessened in this formulation and thus more interpretable (only 3 moderately correlated variables r = 0.34 - 0.52, instead of 9 that ranged from r = 0.13 - 0.84). We replaced *QuestionnaireScore* in the model formula described previously with three psychiatric factors (*AD, CIT, SW*) entered as z-scored fixed effect predictors. The model was: Action ~ Confidence * (AD + CIT + SW + Age + IQ + Gender) + (1 + Confidence | Subject).

Action and confidence. To analyse the basic relationship between task-related variables and psychiatric factors, the analysis approach was the same, but simpler. Dependent variables were: 1) *Accuracy* (outcome of trial *t*, coded as 1 (hit) or 0 (miss), 2) size of bucket updates (*Action*) and 3) reported confidence (*Confidence*). The models were simply: Task Variable ~ AD + CIT + SW + Age + IQ + Gender + (1 | Subject).

Computation model describing behaviour dynamics. To employ model-based analysis, we calculated task prediction error (PE: distance between the particle landing location and the centre of the bucket) and human learning rate (LR^h: change in chosen bucket position from *t* to *t+1* divided by PE on trial *t*) for each trial. Trials where LR^h exceeded the 99th percentile (LR^h > 7.75) and PE = 0 are thought to be unrelated to error-driven learning(Nassar et al., 2016), and were thus excluded from analyses with the model parameters (3.05% of total trials).

In the behavioural task, participants were required to learn the mean of the underlying generative distribution in order to position their bucket at where they hope to catch the greatest number of particles. Their belief on where the particle landing distribution mean could be guided by 1) information gained from the most recent outcome (i.e. moving the bucket with every small shift in particle location), 2) surprising large changes signalling a change in mean distribution (i.e. change-points) and 3) their uncertainty of the distribution mean based on particle landing location experience over trials. To separate these contributions, a quasi-optimal Bayesian computational learning model was used to estimate these parameters thought to underlie task dynamics with MATLAB R2018a (The MathWorks, Natick, MA) using functions from Vaghi *et al.* (Vaghi et al., 2017) (see online supplement for further model details). This

included PE^b (model prediction error, an index of recent outcomes), *CPP* (probability that a trial was a change-point, a measure representing the belief of a surprising outcome) and *RU* (relative uncertainty, the uncertainty owing to the imprecise estimation of the distribution mean; labelled as *(1-CPP)*(1-MC)* in Vaghi *et al.* (Vaghi et al., 2017)). These parameters (where PE^b is taken as its absolute) together with a *Hit* categorical predictor (previous trial was a hit or miss) were used to regress participant adjustments against the benchmark Bayesian model to investigate participant adjustments in reported confidence (*Confidence*; z-scored confidence level on trial *t*) and bucket movements (*Action*) according to the particle landing locations experienced.

Influence of parameters on action and confidence. For the regression on *Action*, following Vaghi *et al.* (Vaghi et al., 2017) and prior literature (McGuire et al., 2014; Nassar et al., 2016; Nassar et al., 2010), all predictors except PE^b were implemented as interaction terms with PE^b. For *Confidence*, we used a similar regression model but without the interaction term with PE^b and with the regressand and predictors z-scored at participant level. Regressions were constructed as mixed-effect models controlled for age, IQ and gender, with the interaction term and main effect of regressors as random effects. The model syntax was written as: Dependent Variable ~ (PE^b + CPP + RU + Hit)*(Age + IQ + Gender) + (1 + PE^b + CPP + RU + Hit | Subject). We obtained similar regression estimates with Vaghi *et al.* (Vaghi et al., 2017), suggesting that action/confidence is appropriately updated with these parameters describing belief updating (Table S1). The main effects of the four predictors were correlated with CIT severity, where Pearson's correlation was used to measure the association between symptom dimension severity and the influence of the learning parameters on action update/confidence. To include all three psychiatric factors in the same analysis model (Figure S5), we entered the transdiagnostic factors as additional z-scored fixed effect predictors into the basic model above, where the equation was: Dependent Variable ~ (PE^b + CPP + RU + Hit)*(AD + CIT + SW + Age + IQ + Gender) + (1 + PE^b + CPP + RU + Hit | Subject). For confidence, a positive interaction between a *Factor score* and PE^b, *CPP, RU* indicates that higher scores on that factor are associated with a decrease in influence of these parameters on confidence. The converse was applicable for significant *Hit*Factor* interactions (as main effect of *Hit* on *Confidence* is opposite signed). For action, as main effect of the parameters on *Action* is inverse

from the main effects on *Confidence*, significant parameter**Factor* interactions are interpreted in reverse.

The code and data to reproduce the main figures of the paper are freely available at https://osf.io/2z6tw/.

Author Contributions

TS and CG conceived of and designed the study and analysis plan. TS coded the experiment, collected and analysed the data. TS and CG wrote the paper.

Acknowledgements

Tricia Seow is supported by a Postgraduate Ussher fellowship from Trinity College Dublin. Claire M. Gillan is supported by a fellowship from MQ: transforming mental health (MQ16IP13). We thank Benedetto de Martino for his helpful comments on the manuscript.

Competing Interests

The authors declare no financial and non-financial competing interests.

Supplementary Material & Figures

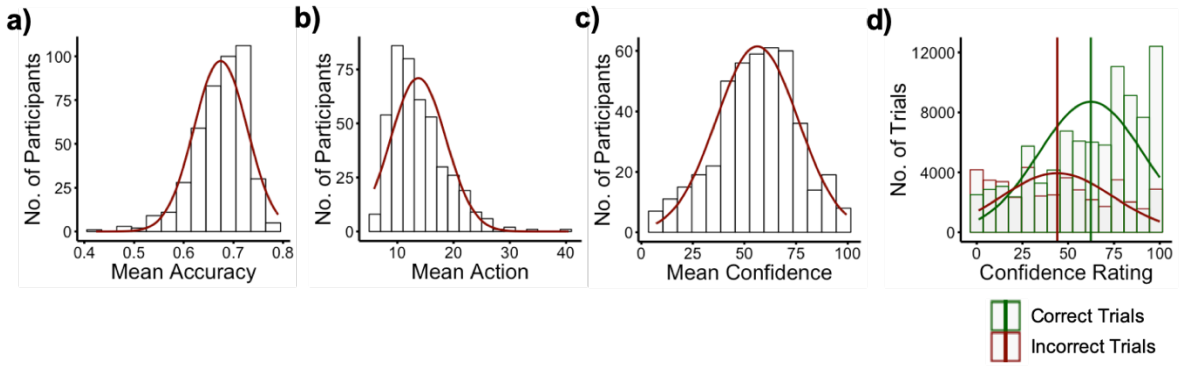

Figure S1. *Behavioural results. Across participants, the distribution of:*
(a) *Mean accuracy.*
(b) *Mean action (the tendency to move the bucket).*
(c) *Mean confidence level.*
(d) *Confidence ratings for correct (green) and incorrect (red) trials. Vertical lines denote mean confidence level for respective trial type.*

Across participants, mean accuracy ranged from 42.33% to 79.00% (mean = 67.42%, SD = 5.38%; Figure S1a), mean action (tendency to move bucket position) ranged from 5.88 to 40.44 (mean = 13.74, SD = 4.91, Figure S1b) and mean confidence level ranged from 7.21 to 99.39 across participants (mean = 56.19, SD = 19.85; Figure S1c). Performance accuracy accounted for only 1.7% of the variance in confidence levels (between-subject correlation: $r = 0.13$, $p < 0.009$). Participants were using the confidence scale appropriately, giving higher confidence after correct trials (mean = 62.42, SD = 28.53), and lower confidence after incorrect trials (mean = 43.98, SD = 30.45) (Figure S1d).

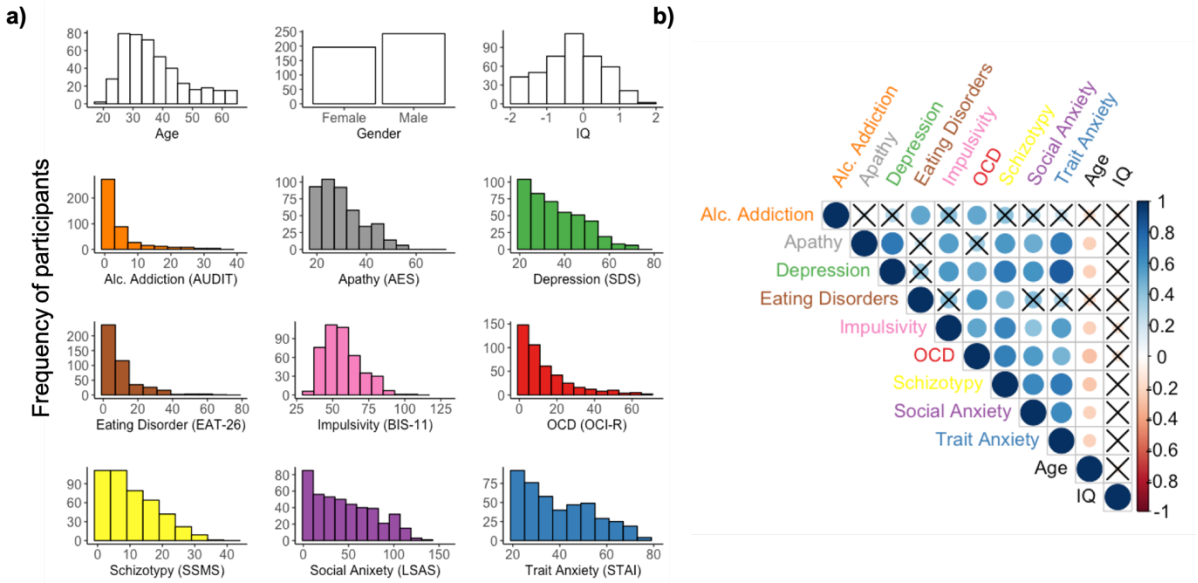

Figure S2. *Demographics and self-reported psychopathology spread.*
(a) *Age, IQ and psychiatric symptoms score distributions across participants.*
(b) *Correlation matrix of mean scores of the nine questionnaires, age and IQ. Colour scale indicates correlation coefficient, size of colour patch indicates significance. X denotes correlation fails 95% Confidence Interval.*

696 **Table S1. Effects of Bayesian Model Parameters on Action and Confidence.**

Predictor	***β (SE)***	***t-value***	***p-value***
	Regression on Action		
PE[b]	*0.33 (0.02)*	*11.61*	*< 0.001 ****
CPP	*0.46 (0.02)*	*20.06*	*< 0.001 ****
RU	*1.37 (0.08)*	*17.25*	*< 0.001 ****
Hit	*-0.77 (0.02)*	*-40.54*	*< 0.001 ****
	Regression on Confidence		
PE[b]	*-0.04 (0.01)*	*-3.57*	*< 0.001 ****
CPP	*-0.20 (0.02)*	*-12.16*	*< 0.001 ****
RU	*-0.24 (0.01)*	*-17.15*	*< 0.001 ****
Hit	*0.26 (0.01)*	*22.84*	*< 0.001 ****

697
698
699
700

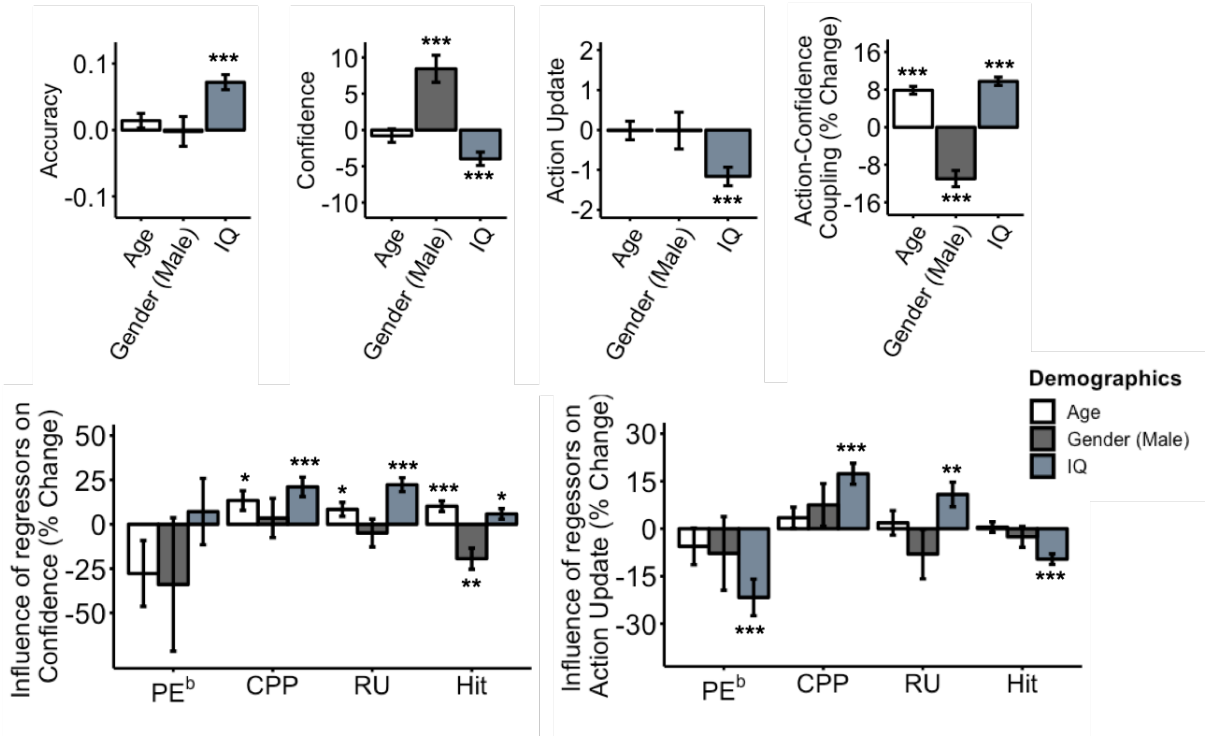

Figure S3. *Associations between age, gender and IQ with accuracy, reported confidence, action update, action-confidence coupling or the influence of the model predictors (PE^B, CPP, RU) and Hit on confidence/action update. Error bars denote standard errors. The Y-axes indicates the change/percentage change in each dependent variable as a function of 1 standard deviation increase of demographic scores. *p < 0.05, **p < 0.01, ***p < 0.001.*

We tested in an exploratory fashion for relationships of task accuracy, action and confidence with age, IQ and gender (Figure S3). IQ was found to predict better performance (β = 0.07, *SE* = 0.01, p < 0.001), lower action updating, (β = -1.16, SE = 0.23, p < 0.001) and lower confidence (β = -3.97, *SE* = 0.92, p < 0.001). Additionally, gender (male) was associated with higher confidence (β = 8.43, *SE* = 1.85, p < 0.001).

IQ, age and gender were controlled for in all analyses. Increased action-confidence coupling was associated to age (β = -6.70, *SE* = 0.08, p < 0.001), and IQ (β = -8.87, *SE* = 0.08, p < 0.001) while decreased in males (β = 0.97, *SE* = 0.15, p < 0.001). For the model-based trial-wise analyses, age was related to an increased influence of CPP (β = -0.03, *SE* = 0.01, p = 0.02), RU (β = -0.02, *SE* = 0.01, p = 0.03) and Hit (β = -0.02, *SE* = 0.01, p = 0.03) on confidence. Males were associated to an increased influence of Hit (β = -0.05, *SE* = 0.02, p = 0.001) on confidence, while IQ predicted increased influence of CPP, RU (CPP: β = -0.05, *SE* = 0.01, p < 0.001, RU: β = -0.05, *SE* = 0.01, p < 0.001) and Hit (β = 0.02, *SE* = 0.01, p = 0.05) on confidence. For action update, only IQ effects were significant – it was related to an increase in CPP (β = 0.08, *SE* = 0.02, p < 0.001) and RU (β = 0.15, *SE* = 0.05, p = 0.006) influence,

725 and decreased PE^b (β = -0.07, *SE* = 0.02, $p < 0.001$) and Hit (β = 0.07, *SE* = 0.01, p
726 < 0.001) influence on action update.
727

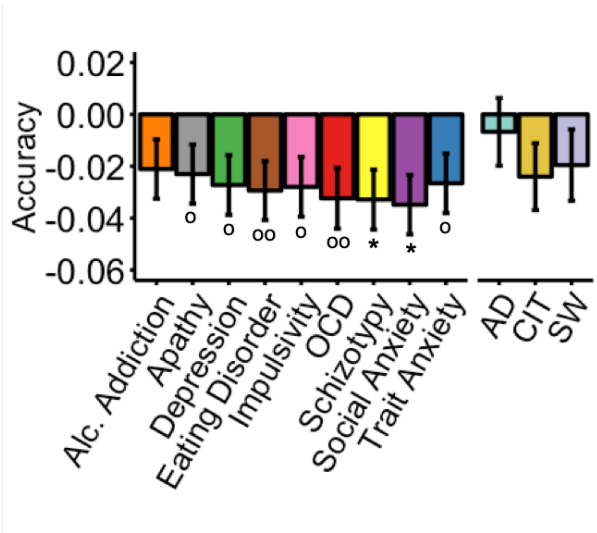

Figure S4. *Associations between accuracy (hit (1) or miss (0)) with psychiatric symptoms or transdiagnostic factors, controlled for age, IQ and gender. Error bars denote standard errors. The Y-axis indicates the change in accuracy as a function of 1 standard deviation of symptom/factor scores.* $^{o}p < 0.05$, $^{oo}p < 0.01$ *uncorrected,* $*p < 0.05$. *Results are Bonferroni corrected for multiple comparisons over number of symptoms/factors.*

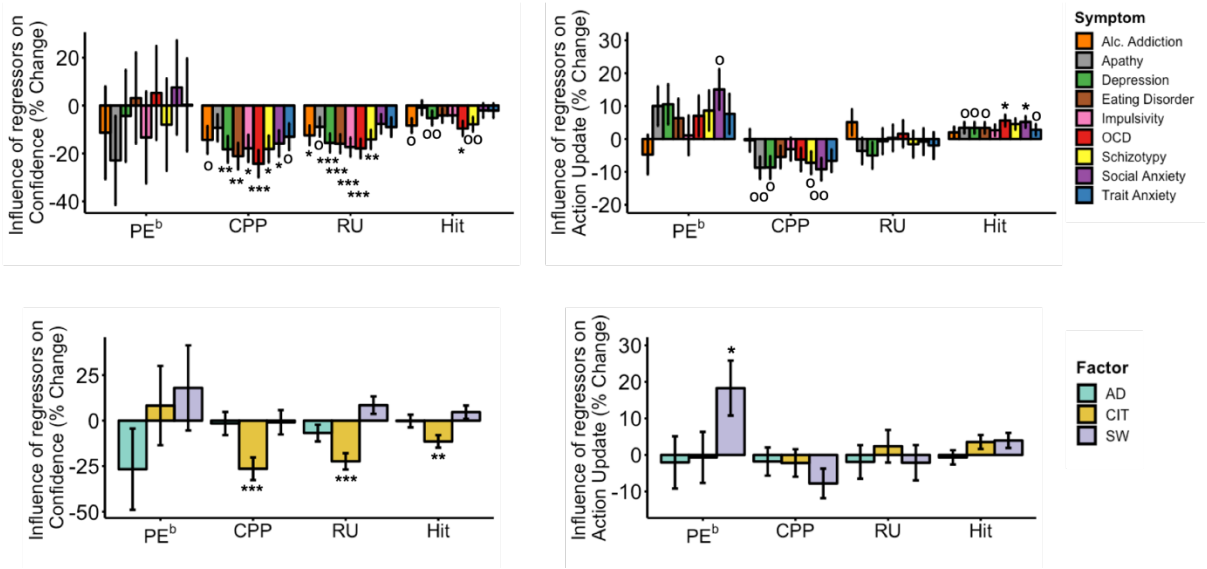

Figure S5. *Regression analyses of trial-wise confidence and action adjustments with psychiatric symptoms/factors, controlled for age, IQ and gender. Each psychiatric symptom was examined in a separate regression, whereas factors were included in the same model. Error bars denote standard errors. The Y-axes indicate the percentage change in each dependent variable as a function of 1 standard deviation of symptom/factor scores. $^{o}p < 0.05$, $^{oo}p < 0.01$ uncorrected, $*p < 0.05$, $***p < 0.001$. Results are Bonferroni corrected for multiple comparisons over number of symptoms/factors.*

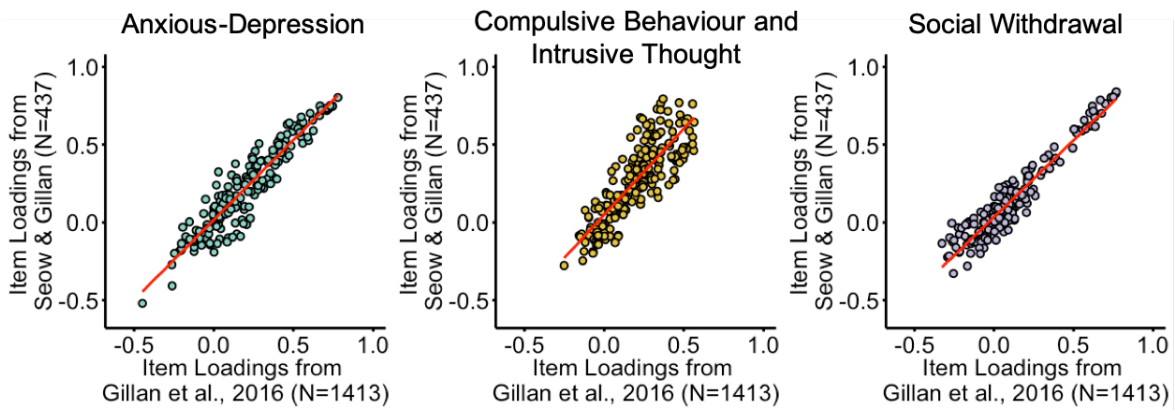

***Figure S6**. Correlations between item loadings obtained from the factor analysis in Gillan et al. (2016) and the present study for each psychiatric symptom dimension. Questionnaire item loadings were highly correlated for all three factors (Anxious-depression: r = 0.94; Compulsive behaviour and intrusive thought: r = 0.85, Social withdrawal: r = 0.95), supporting the reproducibility of the psychiatric symptom dimensions.*

Transdiagnostic symptom dimensions are reproducible. Transdiagnostic factors scores (‘Anxious-depression’, ‘Compulsive behaviour and intrusive thoughts’, ‘Social withdrawal’) in the present study were derived from weights obtained from a prior larger study (N = 1413) (Gillan et al., 2016). This 3-factor structure was previously reproduced in a smaller independent sample (N = 497) (Rouault et al., 2018), and here we again replicated similar psychiatric dimensions with our current data (N = 437) with the factor analysis (Figure S5). For further details of the factor analysis methodology, see Gillan *et al.* (2016) (Gillan et al., 2016).

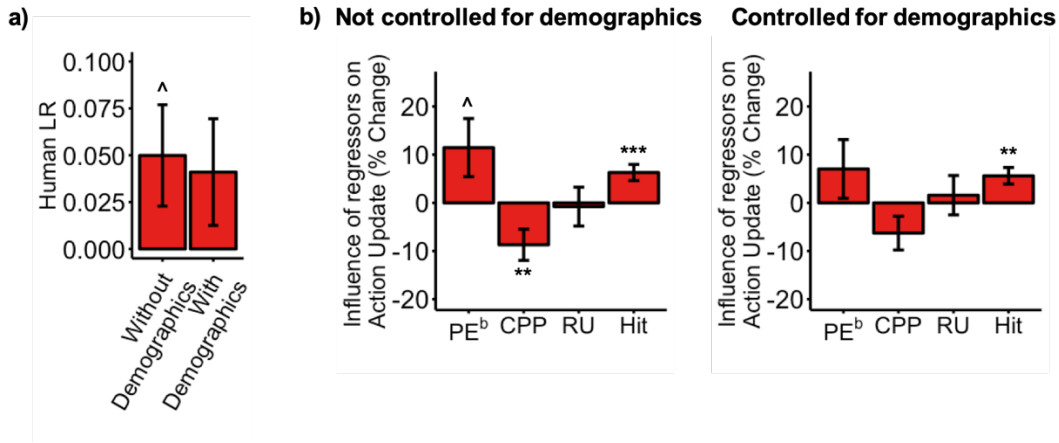

***Figure S7**. Regression analyses of **(a)** human learning rate (ratio of bucket movement and task prediction error) and **(b)** action adjustments in OCD, in a model that controlled for age, IQ and gender and in a model that did not. Error bars denote standard errors. The Y-axes indicate the change/percentage change in dependent variable as a function of 1 standard deviation of OCD symptom scores. ^$p < 0.07$, **$p < 0.01$, ***$p < 0.001$. Results are not Bonferroni corrected for multiple comparisons.*

Action updating effects in OCD with/without controlling for demographics. Vaghi *et al.* (Vaghi et al., 2017) reported that OCD patients exhibited a higher mean learning rate and that their action updates were more strongly influenced by recent information (PE^b) and less to large unexpected environmental changes (CPP). In the course of exploring the source of this discrepancy with our data, we found that when we repeated our analysis without controlling for age, gender and IQ, some of their effects were recovered here. OCD symptoms were associated with changes in learning and sensitivity to both PE^b and CPP in action updating. Specifically, LR^h (β = 0.05, *SE* = 0.03, p = 0.07, uncorr.) and the influence of PE^b on action showed a trend towards a positive association with OCD symptoms (β = 0.04, *SE* = 0.02, p = 0.06, uncorr.) and the influence of CPP on action showed a negative association with OCD symptoms (β = -0.04, *SE* = 0.02, p = 0.007, uncorr.). These discrepancies suggest that demographic characteristics perhaps partially explain the pattern of action updating effects observed in the prior patient study (Figure S7).

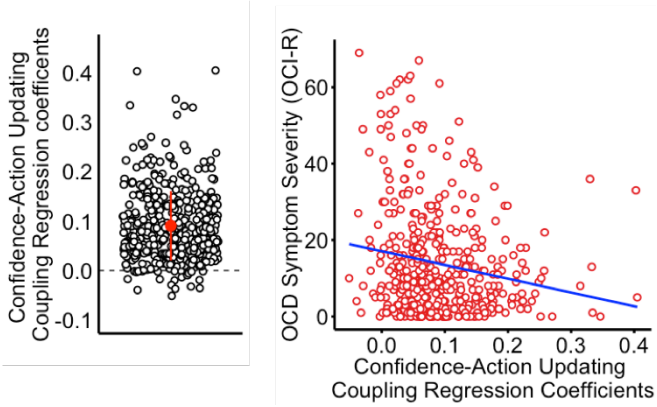

Figure S8. *Regression model where confidence updating was predicted by action updating. Dots represent coefficient estimates for individual participants. Red marker indicates mean and SD. These coefficients were correlated with OCD symptom severity, where confidence-action updating coupling was observed to decrease with increasing OCD symptom severity (r = -0.18, p < 0.001).*

Action-confidence decoupling analysis. Although this has no bearing on our results (or theirs), we note that Vaghi *et al.* (Vaghi et al., 2017) defined action-confidence coupling slightly differently to how we chose to define it in the present paper – they used confidence *updating* (i.e. absolute difference between z-scored confidence from trial *t* and *t*-1), instead of the reported confidence level on trial t. We suggest that z-scored confidence ratings (rather than their change from trial to trial) are more appropriate because this accounts better for instances where a person has several relatively large PEs in a row (as they figure out where to place the bucket), and should thus not rationally 'change' their confidence rating in response to these PEs, but maintain it at a low level. Although we flag this for the interested reader, we underscore that the two measures are correlated and indeed when we use their definition, we similarly show that self-reported OCD symptom severity predicts confidence-action updating decoupling (r = -0.17, t = -3.58, p < 0.001, Figure S8).

Quasi-optimal Bayesian Computational Model

A learning model was used to determine how participants' beliefs of the mean of the particle distribution evolved over time. This was estimated by fitting a reduced quasi-optimal Bayesian learner, in accordance with prior literature (McGuire et al., 2014; Nassar et al., 2016; Nassar et al., 2010) (also see Vaghi *et al.* (Vaghi et al., 2017)), to particle landing location data. Briefly, the model approximates optimal inference by tracking factors that should drive learning: 1) new information gained, 2) surprise from unexpected outcomes and 3) uncertainty of their belief. Using a delta-rule, the model updates its estimate of belief about the particle landing location distribution:

$$B_{t+1} = B_t + \alpha_t \times \delta_t$$

B is the new belief estimate on each trial *t*, which is equal to a point estimation of the mean of the Gaussian distribution where particle locations were sampled (i.e. 1 to 360). Its update is dependent on the learning rate α (LR^b) and model prediction error δ (PE^b). PE^b is calculated as the difference between the belief estimate B_t and the new particle landing location X_t and is a measure of information gained from the most recent trial.

$$\delta_t = X_t - B_t$$

As with common reinforcement learning models, LR^b determines how much new information (PE^b) will update the belief estimate. However, LR^b is dynamic in the current model i.e. can change on every trial. If $LR^b = 0$, new evidence has had no impact on the update of the belief estimate, while $LR^b = 1$ suggests that the new belief estimate is entirely determined by the most recent outcome. The magnitude of LR^b is dependent the statistics of environment with the equation:

$$\alpha_t = \Omega_t + (1 - \Omega_t)\,(1 - \nu_t)$$

The first term, the change-point probability Ω (CPP), represents an estimate of how likely a change in particle location distribution mean has occurred on a given trial. The second term, model confidence ν (MC), represents the uncertainty due to an inaccurate estimation of the mean. For regression analyses, $(1 - \Omega)\,(1 - \nu)$ was

labelled as RU (as the additive inverse of MC is relative uncertainty). These two components allow the model to appropriately update belief according to (i) unexpected changes in the environment (change-points) and (ii) the uncertainty about the distribution mean - thus informing when to disregard outliers when the mean is certain. New outcomes are more influential when the model believes that the distribution mean has changed (i.e. CPP is large) or is less sure about the true distribution mean (i.e. MC is small).

The model generates CPP as the relative likelihood that a new particle location is sampled from a new distribution during a change-point (mean determined by a uniform distribution *U* over all 360 possible locations) or drawn from the same Gaussian (*N*) where the current belief estimate B_t is centered upon. These are influenced by the hazard rate *H*, the probability that the mean of the distribution has changed. We set *H* equal to the hazard rates of the task trials (H = 0.025 or H = 0.125, depending on the trial condition). When the probability of the new particle location coming from a new distribution is high, CPP will be close to 1.

$$\Omega_t = \frac{U(X_t \mid 1{,}360)H}{U(X_t \mid 1{,}360)H + N(X_t \mid B_t, \sigma_t^2)(1-H)}$$

σ_t^2 is the estimated variance of the predictive distribution, which consists of the variance of the generative Gaussian distribution σ_N^2 and the generative variance modulated by MC (ν). As the predictive distribution variance is dependent on MC, it is larger than the generative variance where MC is the smallest (i.e. after change-points, where uncertainty of the new distribution mean is the highest) and will slowly reduce towards the generative variance. Thus, the model describes particle locations occurring after a change-point as less likely sampled from another new distribution.

$$\sigma_t^2 = \sigma_N^2 + \frac{(1-\nu_t)\sigma_N^2}{\nu_t}$$

Lastly, MC is computed for the subsequent trial with a weighted average of the generative variance conditional on a change-point (first term), generative variance conditional on no change-point (second term), and variance due to the model's

uncertainty of whether a change-point occurred (third term) in the numerator. The denominator includes the same terms plus just the generative distribution variance (σ_N^2) representing the uncertainty owing to noise. The full equation is as follows:

$$\nu_{t+1} = \frac{\Omega_t \sigma_N^2 + (1-\Omega_t)(1-\nu_t)\sigma_N^2 + \Omega_t(1-\Omega_t)(\delta_t \nu_t)^2}{\Omega_t \sigma_N^2 + (1-\Omega_t)(1-\nu_t)\sigma_N^2 + \Omega_t(1-\Omega_t)(\delta_t \nu_t)^2 + \sigma_N^2}$$

References

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci, 10*(9), 1214-1221. doi:10.1038/nn1954

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-General Enhancements of Metacognitive Ability Through Adaptive Training. *Journal of Experimental Psychology-General, 148*(1), 51-64. doi:10.1037/xge0000505

Ersche, K. D., Gillan, C. M., Jones, P. S., Williams, G. B., Luitjen, M., de Wit, S., . . . Robbins, T. W. (2016). Carrots and sticks fail to change behavior in cocaine addiction. *Science, 352*(6292), 1468-1471. doi:10.1126/science.aaf3700

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat Neurosci, 8*(11), 1481-1489. doi:10.1038/nn1579

Fineberg, N. A., Potenza, M. N., Chamberlain, S. R., Berlin, H. A., Menzies, L., Bechara, A., . . . Hollander, E. (2010). Probing Compulsive and Impulsive Behaviors, from Animal Models to Endophenotypes: A Narrative Review. *Neuropsychopharmacology, 35*, 591–604. doi:10.1038/npp.2009.185

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol Rev, 124*(1), 91-114. doi:10.1037/rev0000045

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment, 14*(4), 485-496. doi:10.1037//1040-3590.14.4.485

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry, 1*(2), 148-158. doi:10.1016/S2215-0366(14)70275-5

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The Eating Attitudes Test - Psychometric Features and Clinical Correlates. *Psychological Medicine, 12*(4), 871-878. doi:10.1017/S0033291700049163

Gillan, C. M., Apergis-Schoute, A. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Fineberg, N. A., . . . Robbins, T. W. (2015). Functional Neuroimaging of Avoidance Habits in Obsessive-Compulsive Disorder. *American Journal of Psychiatry, 172*(3), 284-293. doi:10.1176/appi.ajp.2014.14040525

Gillan, C. M., & Daw, N. D. (2016). Taking Psychiatry Research Online. *Neuron, 91*(1), 19-23. doi:10.1016/j.neuron.2016.06.002

Gillan, C. M., Fineberg, N. A., & Robbins, T. W. (2017). A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychological Medicine, 47*(9), 1528-1548. doi:10.1017/S0033291716002786

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife, 5*. doi:10.7554/eLife.11305

Gillan, C. M., Morein-Zamir, S., Kaser, M., Fineberg, N. A., Sule, A., Sahakian, B. J., . . . Robbins, T. W. (2014). Counterfactual Processing of Economic Action-Outcome Alternatives in Obsessive-Compulsive Disorder: Further Evidence of Impaired Goal-Directed Behavior. *Biological Psychiatry, 75*(8), 639-646. doi:10.1016/j.biopsych.2013.01.018

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., . . . Robbins, T. W. (2014). Enhanced Avoidance Habits in Obsessive-

Compulsive Disorder. *Biological Psychiatry, 75*(8), 631-638. Rdoi:10.1016/j.biopsych.2013.02.002

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cogn Affect Behav Neurosci*. doi:10.3758/s13415-015-0347-6

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the Balance Between Goal-Directed Behavior and Habit Learning in Obsessive-Compulsive Disorder. *American Journal of Psychiatry, 168*(7), 718-726. doi:10.1176/appi.ajp.2011.10071062

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive-compulsive disorder. *Philos Trans R Soc Lond B Biol Sci, 369*(1655). doi:10.1098/rstb.2013.0475

Hauser, T. U., Allen, M., Consortium, N., Rees, G., & Dolan, R. J. (2017). Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. *Sci Rep, 7*(1), 6614. doi:10.1038/s41598-017-06116-z

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience, 19*(3), 404-413. doi:10.1038/nn.4238

Huys, Q. J. M., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks, 24*(6), 544-551. doi:10.1016/j.neunet.2011.03.001

Hyman, S. E. (2007). Opinion - Can neuroscience be integrated into the DSM-V? *Nature Reviews Neuroscience, 8*(9), 725-U716. doi:10.1038/nrn2218

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry, 167*(7), 748-751. doi:10.1176/appi.ajp.2010.09091379

Liebowitz, M. R. (1987). Delineating Social Phobia. *British Journal of Psychiatry, 150*, 718-719.

Marin, R. S. (1991). Apathy - a Neuropsychiatric Syndrome. *Journal of Neuropsychiatry and Clinical Neurosciences, 3*(3), 243-254.

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research, 78*(2-3), 293-296. doi:10.1016/j.schres.2005.06.020

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron, 84*(4), 870-881. doi:10.1016/j.neuron.2014.10.013

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron, 88*(1), 78-92. doi:10.1016/j.neuron.2015.09.039

Nassar, M. R., Bruckner, R., Gold, J. I., Li, S. C., Heekeren, H. R., & Eppinger, B. (2016). Age differences in learning emerge from an insufficient representation of uncertainty in older adults. *Nat Commun, 7*, 11609. doi:10.1038/ncomms11609

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J Neurosci, 30*(37), 12366-12378. doi:10.1523/JNEUROSCI.0822-10.2010

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51*(6), 768-774. doi:10.1002/1097-4679(199511)51:6<768::Aid-Jclp2270510607>3.0.Co;2-1

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). Incentives Boost Model-Based Control Across a Range of Severity on Several Psychiatric Constructs. *Biol Psychiatry, 85*(5), 425-433. doi:10.1016/j.biopsych.2018.06.018

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48. doi:10.1006/cogp.1999.0735

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biol Psychiatry, 84*(6), 443-451. doi:10.1016/j.biopsych.2017.12.017

Ruscio, A. M., Stein, D. J., Chiu, W. T., & Kessler, R. C. (2010). The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Molecular Psychiatry, 15*(1), 53-63. doi:10.1038/mp.2008.94

Saunders, J. B., Aasland, O. G., Babor, T. F., Delafuente, J. R., & Grant, M. (1993). Development of the Alcohol-Use Disorders Identification Test (Audit) - Who Collaborative Project on Early Detection of Persons with Harmful Alcohol-Consumption .2. *Addiction, 88*(6), 791-804. doi:10.1111/j.1360-0443.1993.tb02093.x

Snorrason, I., Lee, H. J., de Wit, S., & Woods, D. W. (2016). Are nonclinical obsessive-compulsive symptoms associated with bias toward habits? *Psychiatry Res, 241*, 221-223. doi:10.1016/j.psychres.2016.04.067

Spielberger, C. D., Gorsuch, R.L., Lushene, R.D., Vagg, P.R., Jacobs, G.A. (1983). *Manual for the state-trait anxiety inventory (STAI)* Palo Alto, CA, USA: Consulting Psychologists Press.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology, 25*, 85-92. doi:10.1016/j.conb.2013.12.007

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron, 96*(2), 348-354 e344. doi:10.1016/j.neuron.2017.09.006

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., . . . Bullmore, E. T. (2014). Disorders of compulsivity: a common bias towards learning habits. *Mol Psychiatry*. doi:10.1038/mp.2014.44

Voon, V., Derbyshire, K., Ruck, C., Irvine, M. A., Worbe, Y., Enander, J., . . . Bullmore, E. T. (2015). Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry, 20*(3), 345-352. doi:10.1038/mp.2014.44

Zung, W. W. K. (1965). A Self-Rating Depression Scale. *Archives of General Psychiatry, 12*(1), 63-70.