NOVA MEDICAL SCHOOL FACULDADE DE CIÊNCIAS MÉDICAS

UNIVERSIDADE NOVA DE LISBOA

KNOWLEDGE VERSUS EXPERIENCE

Exploring model-based and model-free reinforcement learning in obsessive-compulsive disorder

Pedro Manuel Ferreira de Castro Rodrigues

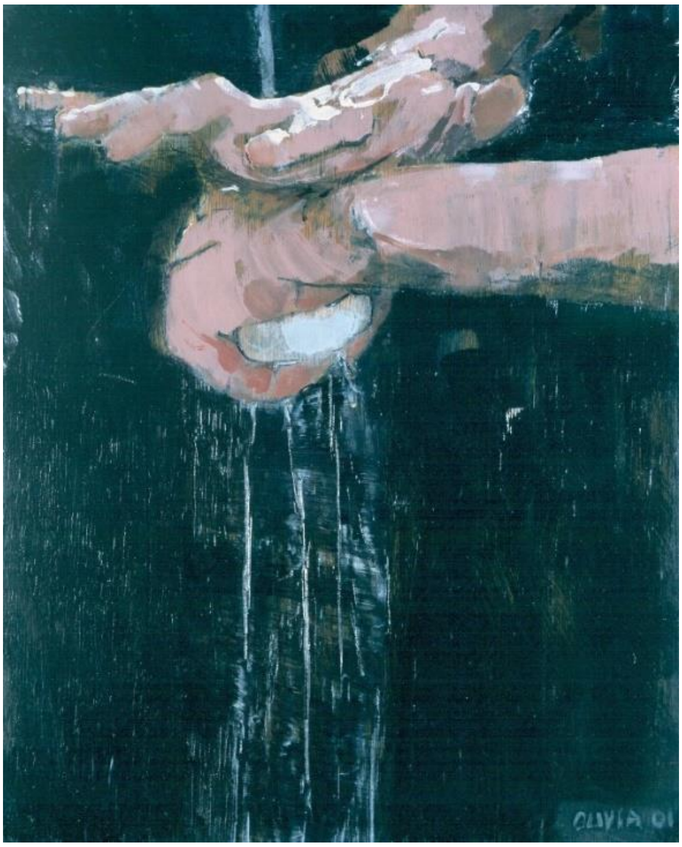
A thesis submitted in partial fulfilment of the requirements for the Doctoral Degree in Medicine

September 2019

Knowledge versus experience

Exploring model-based and model-free reinforcement learning in obsessive-compulsive disorder

Pedro Manuel Ferreira de Castro Rodrigues



Olivia Gillow, 'Insistent rub', 2001

A thesis submitted in partial fulfilment of the requirements for the
Doctoral Degree in Medicine
Specialization: Health of Populations

NOVA Medical School | Faculdade de Ciências Médicas (NMS | FCM),

Universidade Nova de Lisboa

Supervisors:

Albino J. Oliveira-Maia, Assistant Professor, NMS | FCM

J. Bernardo Barahona-Corrêa, Assistant Professor, NMS | FCM

This work was developed in the context of the Doctoral Programme in Medicine (6th edition) of NOVA Medical School, Nova University of Lisbon, Lisbon, Portugal, under the supervision of Professors Albino Oliveira-Maia and Bernardo Barahona-Corrêa.

The experimental work was carried out and approved by the Ethics Committees of Champalimaud Centre for the Unknown, Centro Hospitalar Psiquiátrico de Lisboa, New York State Psychiatric Institute and Zuckerman Mind Brain Behavior Institute (Columbia University).

The candidate was supported by a fellowship from the Foundation for Science and Technology (Portugal; SFRH/SINTD/94350/2013) and by a Fulbright Research Grant from the Bureau of Educational and Cultural Affairs of the United States Department of State

Acknowledgements

To my mother, Ana Luísa, for her unconditional love and for teaching me the value of knowledge and of taking care of others,

To my father, Manuel, for his unconditional love and for teaching me about curiosity, creativity and passion,

To my sister Sara, for growing up with me, and to my sister Maria, for bringing so much joy,

To my grandmother Manuela, for her patience, and my grandfather Staline, for his courage,

To my grandmother Luísa, for her laughs, and my grandfather Hélder, for his wisdom,

To Albino Oliveira-Maia, for being my mentor and my friend, for the constant challenges (this PhD being the main one), for the insightful discussions and for guiding me when things were good and when things were less good,

To Bernardo Barahona-Corrêa, for sharing his deep knowledge about OCD, for helping me to think about the clinical implications of research and for teaching me to present complex concepts in a clearer way,

To Rui Costa, for being the scientist that made me want to do science,

To Thomas Akam, for his computationally incredible model-based system,

To Marta Camacho, for her friendship, intelligence, persistence and wit,

To Vitor Paixão, for his Matlab skills and his passion for Sport Lisboa e Benfica,

To Ana Fernandes, for making me believe that this was possible and for being there whenever I needed help,

To Joaquim Alves da Silva, for inviting me to resume my teaching activity and for watching me present the same thing four times in a row,

To Gabriela Ribeiro, for all the advice and all the coffees,

To José Oliveira, for being a role model on how a psychiatry resident can make good research,

To the rest of the Neuropsychiatry Unit at Champalimaud (especially Jaime, Sílvia, Gonçalo and Ana) for all the feedback, interest and support,

To Helen Blair Simpson, for carrying this project to the New York State Psychiatric Institute,

To Ivar and Yael, for collecting an incredible amount of clinical data,

To both branches of the Costa Lab, for making research fun,

To Gabriela Martins, for turning abstract ideas into concrete events,

To Mariana Lopes Correia, for always finding space in someone's schedule and time for a laugh,

To Rodrigo Oliveira, Nuno Loureiro and Michael Pereira, for all the questions and suggestions,

To Akram Bakkour, for helping me to design my first fMRI task,

To Suwon and Dania, for being so fast and competent in collecting imaging data,

To Akira, Vivek, Ines and the rest of the basketball crew in NY,

To Cath, Som, Mert, Dennis, André, Nico and the basketball crew in Lisbon,

To Catarina Soares, for her decades of work with OCD patients,

To Alice Nobre, for trusting me and always letting me carry my ideas and research projects in the department she directs,

To Paula Nunes, for teaching me psychiatry,

To Beatriz Navarro, for helping with the aesthetic aspects of my work,

To all my friends, especially João Colaço, Miguel, João Araújo, Sandro, Luís, Joana Colaço and Beatriz Lourenço, for always being there for me and for forgiving all skipped dinners in the last five years,

And to all the volunteers who participated in our studies, particularly the patients with OCD, to whom I hope this work can be helpful someday,

Summary

Obsessive-compulsive disorder (OCD) is a common, chronic and disabling neuropsychiatric condition for which current treatments are ineffective in a large proportion of cases. The gold-standard instrument to assess the severity of OCD symptoms is the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS), which was recently revised (Y-BOCS-II). However, its construct validity has been reported has moderate and its criterion-related validity for the diagnosis of OCD has never been tested. In the first chapter of this dissertation, I tested, for the first time, criterion-related validity of the Y-BOCS-II and demonstrated that a cut-off of 13 (total score) attains the best balance between sensitivity and specificity for the diagnosis of OCD. However, I confirmed that its divergent validity is far from excellent. This last finding led me to search for other potential markers of OCD.

Several abnormalities have been demonstrated in OCD patients in studies using neuropsychological and neuroimaging approaches, but we still lack a consistent marker for the disorder which is able to discriminate patients with OCD from healthy subjects or from patients with other mental disorders, which is sensitive to treatment-induced changes, and which can be mapped to brain circuits or function. An approach which has been followed over the last decade is considering OCD as a disorder of action learning systems of the brain. Sequential decision tasks have recently emerged as an influential and sophisticated tool to investigate action learning in humans through the reinforcement learning (RL) framework. According to the RL framework, actions can be learned in two different ways: model-based control works by learning a model of the dynamics of the environment and later using that model to plan future behavioral trajectories, while model-free control works by storing the estimated value of recently taken actions and updating these estimates by trial-and-error. Sequential decision tasks have been used to assess associations between dysfunction in RL control systems and certain behavioral disorders, such as OCD, where an unbalance between model-based and model-free RL has been hypothesized. In fact, using the most commonly applied sequential decision task, the two-step task, evidence has been produced suggesting that OCD patients have a deficit in model-based learning. However, in this specific paradigm, subjects typically receive detailed information about task structure prior to performing the task. Thus, it remains unclear how different RL systems contribute when subjects learn exclusively from experience, and how explicit information about task structure modifies RL strategy. To address these questions, I created a sequential decision task requiring minimal prior instruction, the reduced two-step task. I assessed performance both prior to and after delivering explicit information on task structure, in healthy volunteers, patients with OCD and patients with other mood and anxiety disorders. Initially model-free control dominated, with model-based control emerging only in a minority of subjects after significant task experience, and not at all in patients with OCD, who had instead a tendency to increase their use of model-free control. Once explicit information about task structure was provided, a dramatic increase in the use of model-based RL was observed,

similarly across healthy volunteers and both patient groups, including OCD. The debriefing also significantly decreased the use of model-free RL in healthy volunteers and in patients with mood and anxiety disorders, but not in OCD patients. Additionally, after instructions, model-free action value updates were influenced more by state values and less by trial outcomes, in all groups, and subject choices became more perseverative in healthy subjects, consistent with changes in exploration strategy. These results help in clarifying the RL profile for patients with OCD, with unspecific findings of deficient model-based control, and more specific findings of enhanced model-free control, in both cases prior to information about task structure.

Finally, as the literature is not yet consensual on how model-free and model-based RL systems interact in human brain circuits, I developed a functional magnetic resonance imaging (fMRI) protocol to assess uninstructed and instructed sequential action choice. Preliminary results in healthy subjects suggest that the fMRI version of the reduced two-step task allows to separate predominantly model-free control (before instructions) from predominantly model-based control (after instructions), in the same subject, task structure and environment. Across all sessions, choice events were associated with increases blood-oxygen-level-dependent (BOLD) activity in the left precentral gyrus and reward events were associated with increased BOLD activity in the ventral striatum. I found that explicit knowledge about task structure modifies blood-oxygen-level-dependent (BOLD) activity in the paracingulate cortex (medial prefrontal cortex) during the transition from the first- to the second-step of the task. Future directions include using multivariate pattern analysis techniques to explore how the brain represents state space in sequential decision tasks and applying the current fMRI protocol in clinical populations.

Resumo

A Perturbação Obsessivo-Compulsiva (POC) é uma doença neuropsiquiátrica comum, grave e incapacitante, para a qual os tratamentos actuais são ineficazes num grande número de casos. O instrumento mais utilizado para avaliar a gravidade de sintomas obsessivo-compulsivos é a *Yale-Brown Obsessive-Compulsive Scale* (Y-BOCS), que foi recentemente revista (Y-BOCS-II). No entanto, a sua validade de construto (tanto divergente como convergente) tem sido reportada como moderada e a sua validade de critério para diagnóstico de POC nunca foi testada. No primeiro capítulo desta tese testei, pela primeira vez, a validade de critério da Y-BOCS-II e demonstrei que um ponto de corte de 13 (pontuação total) atinge o melhor balanço entre sensibilidade e especificidade para o diagnóstico de POC. No entanto, confirmei que a sua validade divergente está longe de ser excelente. Este último achado levou-me a procurar outros potenciais marcadores de POC.

Têm sido demonstradas várias anomalias em doentes com POC utilizando tarefas neuropsicológicas ou técnicas de neuroimagem. Contudo, não existe ainda um marcador consistente para esta perturbação, que seja capaz de discriminar eficazmente pacientes que sofrem de POC, que seja sensível à mudança após intervenções terapêuticas e para o qual seja possível estabelecer uma correspondência com circuitos ou função cerebral. Uma abordagem que tem sido seguida nos últimos anos considera a POC como sendo caracterizada por uma disfunção nos sistemas cerebrais responsáveis pela aprendizagem de acções. As tarefas de decisão sequencial emergiram recentemente como um instrumento importante e sofisticado para estudar a aprendizagem de acções em humanos através da abordagem de *reinforcement learning (*RL*)*. De acordo com a teoria subjacente ao RL, as acções podem ser aprendidas de duas formas distintas: um sistema *model-based* funciona através da construção de um modelo interno das dinâmicas do ambiente e utiliza esse modelo para planear trajectórias comportamentais futuras, por oposição a um sistema *model-free*, que funciona armazenando o valor estimado das acções que foram implementadas recentemente e actualizando essas estimativas por tentativa e erro. As chamadas tarefas de decisão sequencial têm vindo a ser utilizadas para estabelecer associações entre disfunção de sistemas cerebrais de RL e algumas perturbações neuropsiquiátricas, como a POC, sendo que um desequilíbrio entre os sistemas *model-based* e *model-free* tem sido descrito. Através da aplicação de uma dessas tarefas de decisão sequencial, a *two-step* task, existe evidência que sugere que os doentes com POC têm um défice no sistema *model-based.* No entanto, neste paradigma em particular, antes de desempenhar esta tarefa os indivíduos recebem informação detalhada sobre a estrutura da mesma. Assim, não é claro como os dois principais sistemas de RL interagem quando os indivíduos aprendem exclusivamente através de interacção com o ambiente e como a informação explícita afecta as estratégias de RL. No segundo capítulo desta tese, desenvolvi uma nova tarefa de decisões sequenciais que permite não só quantificar o uso de estratégias *model-based* RL e *model-free* RL, mas também diferenciar entre o impacto do conhecimento

explícito da estrutura da tarefa e o impacto da experiência na mesma. Os resultados da aplicação da tarefa em indivíduos saudáveis demonstram que inicialmente a escolha de acções é controlada por aprendizagem *model-free*, com a aprendizagem *model-based* emergindo apenas numa minoria de indivíduos depois de experiência significativa com a tarefa, não emergindo de todo em indivíduos com POC, que por sua vez mostraram tendência para aumentar o uso de *model-free* RL com a experiência. Quando foi dada informação explícita sobre a estrutura da tarefa, observou-se um aumento dramático do uso de aprendizagem *model-based*, tanto nos voluntários saudáveis como em ambos os grupos clínicos. A informação explícita diminuiu o uso do sistema de aprendizagem *model-free* nos voluntários saudáveis e nos pacientes com perturbação do humor e ansiedade, mas essa diminuição não foi estatisticamente significativa no grupo de doentes com POC. Para além disso, depois das instruções, verificou-se em todos os grupos que a actualização do valor das acções aprendidas através do sistema *model-free* passou a ser mais influenciada pelo valor dos estados atingidos e menos influenciada pela consequência dos ensaios. Outro efeito da informação explícita sobre a estrutura da tarefa nos indivíduos saudáveis foi tornar as escolhas mais perseverantes, o que é consistente com uma modificação da estratégia de exploração. Estes resultados ajudam a clarificar o perfil de utilização de estratégias de RL dos pacientes com POC, que apresentam défice inespecíficos de aprendizagem *model-based* e achados mais específicos de maior uso de aprendizagem *model-free*, em ambos os casos antes de obterem informação sobrea estrutura da tarefa.

Por fim, como a literatura ainda não é consensual sobre a interação entre um eventual sistema de *model-based* RL e um sistema de *model-free* RL nos circuitos cerebrais em humanos, devenvolvi um protocolo de ressonância magnética funcional para avaliar a escolha de ação sequencial com e sem instruções. Os resultados preliminares, em indivíduos saudáveis, sugerem que a *reduced two-step task* permite separar comportamento que utiliza aprendizagem predominantemente *model-free* (antes das instruções) de comportamento que utiliza aprendizagem predominantemente *model-based* (após as instruções), no mesmo indivíduo, estrutura da tarefa e ambiente. A análise dos dados de imagem funcional sugere que o conhecimento explícito sobre a estrutura da tarefa modifica a atividade neuronal no córtex paracingulado (cortex prefrontal medial) durante a transição do primeiro para o segundo passo da tarefa. Objectivos futuros incluem o uso de técnicas de análise multivariada para explorar a representação cerebral dos estados da tarefa e a aplicação deste protocolo de ressonância magnética funcional em populações clínicas.

List of relevant publications

- Castro-Rodrigues P, Akam T, Snorasson I, Camacho M, Paixão V, Barahona-Corrêa JB, Dayan P, Blair Simpson H, Costa RM, Oliveira-Maia AJ. From experience to knowledge in reinforcement learning. (in preparation).

- Castro-Rodrigues P, Camacho M, Almeida S, Marinho M, Soares C, Barahona-Corrêa JB and Oliveira-Maia AJ, Criterion Validity of the Yale-Brown Obsessive-Compulsive Scale Second Edition for Diagnosis of Obsessive-Compulsive Disorder in Adults. *Front. Psychiatry*, 2018, 9:431.

- Castro-Rodrigues P, Oliveira-Maia AJ. "Mecanismos cerebrais de aprendizagem pelo reforço na perturbação obsessivo-compulsiva". *Psilogos* 2017; 15, 1.

- Barahona-Corrêa JB, Camacho, M., Castro-Rodrigues P, Costa R & Oliveira-Maia AJ. From Thought to Action: How the Interplay Between Neuroscience and Phenomenology Changed Our Understanding of Obsessive-Compulsive Disorder. *Frontiers in psychology*, 2015, 6.

- Oliveira-Maia A., Castro-Rodrigues P. Brain-derived neurotrophic factor: a biomarker of obsessive-compulsive disorder? *Frontiers in Neuroscience*, 2014, 9, 134.

- Castro-Rodrigues P, Oliveira-Maia A. "Exploring the effects of depression and treatment of depression in reinforcement learning. *F Integrative Neuroscience*, 2013, 7.

Contents

Figures index

Tables index

Abbreviations list

ACC	Anterior cingulate cortex
AI	Artificial intelligence
AUC	Area under the (ROC) curve
BDI-II	Beck Depression Inventory-II
BOLD	Blood-oxygen level-dependent
CBT	Cognitive-behavioral therapy
COI	Coimbra Obsessive Inventory
DASS	Depression Anxiety Stress Scales
DBS	Deep brain stimulation
DMS	Dorsomedial striatum
DLS	Dorsolateral striatum
DLPFC	Dorsolateral prefrontal cortex
ERP	Exposure and response prevention
fMRI	Functional magnetic resonance imaging
GDB	Goal-directed behavior
GLM	General linear model
GWAS	Genome-wide association study
HB	Habitual behavior
lOFC	Lateral orbitofrontal cortex
IUS	Intolerance to Uncertainty Scale
MA	Mood and anxiety disorders
MB	Model-based
MDP	Markov decision process
MF	Model-free
mOFC	Medial orbitofrontal cortex
mPFC	Medial prefrontal cortex
MVPA	Multi-voxel pattern analysis
OC	Obsessive compulsive (symptoms)

OCD	Obsessive-compulsive disorder
OFC	Orbitofrontal cortex
PET	Positron-emission tomography
PY-BOCS-II	Portuguese translation of the Y-BOCS-II
RI	Random-interval (schedule of reinforcement)
RL	Reinforcement learning
ROC	Receiving operating characteristics (curve)
RPE	Reward prediction error
RR	Random-ratio (schedule of reinforcement)
RSA	Representation similarity analysis
SPE	State prediction error
SRI	Serotonin reuptake inhibitor
SSRI	Selective serotonin reuptake inhibitors
SSRT	Stop-signal reaction time
STAI	State-Trait Anxiety Inventory
TD	Temporal difference (learning)
TMS	Transcranial magnetic stimulation
VI	Variable-interval (schedule of reinforcement)
vmPFC	Ventromedial prefrontal cortex
VR	Variable-ratio (schedule of reinforcement)
VTA	Ventral tegmental area
WM	Working memory
Y-BOCS	Yale-Brown Obsessive-Compulsive Scale
Y-BOCS-II	Yale-Brown Obsessive-Compulsive Scale Second Edition

Chapter 1. Introduction

1.1. Obsessive-compulsive disorder

Obsessive-compulsive disorder (OCD) is a chronic and incapacitating neuropsychiatric disorder characterized by the presence of obsessions or compulsions, most frequently both[1,2]. Obsessions are recurrent and persistent thoughts, urges or images experienced as intrusive, unwanted or inappropriate, that cause marked anxiety or distress. The individual with obsessions makes (often unsuccessful) attempts to ignore or suppress those thoughts, urges or images. Compulsions are repetitive behaviors or mental acts that the subject feels driven to perform repeatedly, frequently in response to an obsession or according to rules that must be applied rigidly. These behaviors or mental acts are aimed at reducing anxiety or distress or at preventing some unwanted event, but are clearly excessive or unrealistically related with what they are supposed to reduce or prevent[1,2].

The lifetime prevalence of OCD is about 3% worldwide, having been estimated at 2.3% in the US and at 5.3% in Portugal[3–5]. It is the fourth most common psychiatric diagnosis – after phobias, substance-related disorders and depression[6]. The incidence of the disorder has a bimodal distribution, with a first peak in childhood (average age of onset of 10 years old)[7] and another peak in adulthood (average age of onset of 20 years old), with earlier onset for males[8]. Approximately one third of adult OCD patients report that their symptoms started during childhood and, among children, two thirds of cases are boys[9,10].

OCD is not only common – it also produces a high level of impairment. It is ranked by the World Health Organization as the tenth medical condition overall in years lived with disability[11]. OCD patients typically spend a large amount of time (usually more than an hour per day) performing compulsions or having obsessions [2]. Almost all patients with OCD report that their obsessions cause them significant distress and anxiety, as opposed to similar intrusive thoughts in persons who do not have OCD[12]. The obsessions and compulsions may make even the simplest of daily activities stressful and time consuming. In terms of quality of life, persons who suffer from OCD report a persistent decrease compared to controls[12]. Moreover, when compared with persons with anxiety or depressive disorders, an OCD patient is less likely to be married, more likely to be unemployed and more likely to report impaired social and occupational functioning[13].

The recommended treatment for OCD is based on a combination of pharmacotherapy and cognitive behavior therapy (CBT). OCD pharmacotherapy typically includes serotonin reuptake inhibitors (SRI’s, mainly clomipramine and selective SRI’s, SSRIs, such as fluvoxamine or paroxetine) according to a dose-response relationship (higher doses typically needed for better clinical response), or SSRI’s combined with antipsychotic agents,[14–17]. The most widely studied form of CBT is a variant called exposure and response prevention (ERP), which consists of a graded, extended exposure to stimuli or situations that typically cause obsessions to occur, integrated with instructions that help the subject to abstain from performing the compulsions. Several studies demonstrate that ERP is effective in reducing OCD symptoms[18–21]. However, controlled trials have shown that even with a combination of ERP and pharmacotherapy of adequate duration and dosage 30 to 50% of patients remain treatment resistant[22]. In a recent naturalistic 2-year prospective study of the course of OCD, only 6% of treatment-seeking adult patients achieved complete remission[23].

1.1.1. Assessment

While obsessions and compulsions form the core phenomenology of OCD, there is considerable heterogeneity in symptom presentation, which can vary widely across individuals. The gold-standard instruments used to assess symptom severity in OCD patients are the Yale-Brown Obsessive-Compulsive Scales (Y-BOCS). The Y-BOCS has been used as the primary outcome in virtually all contemporary clinical trials in OCD[24,25]. The first edition of the Y-BOCS was published in 1989 by Goodman and colleagues[26] and is a clinician-rated instrument divided in two sections: a dichotomous symptom checklist which assesses the presence of several types of obsessions or compulsions and a severity scale which quantifies the impact of the symptoms identified in the checklist. The severity scale is rated based on a semi-structured interview that assesses the severity, frequency, duration and functional impact of obsessions and compulsions separately. Factor analysis using the symptom checklist of the Y-BOCS, has been used to try to identify OCD subtypes. While initial factor studies identified a three-factor model[27], later studies proposed a four-[28,29] or five-factor structure[30]. The largest meta-analytic study that investigated symptom dimensions in OCD found support for a four factor solution: the first factor consisted of symmetry obsessions with repeating, ordering and counting compulsions; the second factor included aggressive, sexual, religious, somatic and harm obsessions accompanied by checking compulsions; the third factor consisted of contamination obsessions with cleaning compulsions and the fourth factor included hoarding obsessions and compulsions[31]. Factor analysis of the severity scale of the Y-BOCS has typically shown a two-factor structure consistent with distinct severity dimensions for obsessions and compulsions[32–35]. However, other studies found a different two-factor structure, comprising an Interference factor (i.e., distress related to obsessions or not performing compulsions, time occupied by obsessive thoughts or compulsions, functional interference due to obsessions or compulsions) and a Resistance/Control factor (efforts to resist obsessions or compulsions, degree of control over obsessions or compulsions) [36,37].

Despite showing consistent reliability, the first edition of the Y-BOCS raised concerns related to its factor structure, its divergent validity and its poor sensitivity to treatment-changes in severe cases, leading to a second edition of the Y-BOCS (Y-

BOCS-II) being published in 2010[38]. The Y-BOCS-II has shown excellent reliability, good convergent validity and better sensitivity to change in severe cases [38]. However, its discriminant validity is limited, criterion-related validity has never been tested, and the factor structure of its symptom checklist remains a matter of debate[38–41]. On the second chapter of the present thesis, I will explore the validity of the Y-BOCS-II.

1.1.2. Psychological theories

Although OCD is a frequent and clinically well-characterized disorder, the pathophysiology underlying OCD remains poorly understood[22]. Our current knowledge on the etiology of OCD comes from various sources and scientific disciplines. The first tentative explanations were psychological formulations. Psychoanalytic theories of OCD lack empirical evidence but there is large evidence-based literature supporting cognitive and behavioral models of OCD. In 1950, Dollard and Miller[42] adapted the two-factor model originally developed by Mowrer for fear conditioning[43]. According to this model, an individual with OCD first learns anxiety or distress form associations between these feelings and an original neutral stimulus[42]. Then, through a process of conditioning, the originally neutral stimulus becomes a conditioned anxiety stimulus to which the patient gradually develops avoidance and escape responses. These responses – through their effectiveness in reducing anxiety – are strengthened and maintained over time. This cycle provided an interesting explanation for the compulsive aspect of OCD but failed to address how obsessions arise. Later cognitive theories proposed that obsessions were formed from a base of cognitive bias such as inflated concerns about normal events and a remarkably high expectation of negative consequences from these otherwise normal events[44]. According to this model, compulsions are considered rational avoidance behaviors that arise as a response to fear, anxiety and irrational beliefs. These cognitivist accounts paved the way for Albert Bandura’s social cognitive learning theory and modern cognitive-behavioral approaches to OCD[45]. According to this model, intrusive thoughts can become obsessions in subjects with OCD because they are appraised as personally important, highly unacceptable or immoral, or posing a threat. These cognitive appraisals will lead to high amounts of distress, which the patient tries to diminish via compulsions that result in temporary anxiety reduction, thus reinforcing the maladaptive beliefs that led to the obsession and perpetuating a circle. The cognitive-behavioral model for OCD has received the most empirical support of any psychological theory, both in terms of experimental evidence and through the use of the psychotherapy practice which shares the same name[46].

Although psychological theories provide comprehensive models for OCD symptoms *per se*, they do not explain why and how they arise in the brain. OCD is a clinically heterogenous phenotype and, as a consequence, biological markers are needed in order to establish more homogenous samples[47]. With that in mind, we turn to what genetics, neuroimaging and neuropsychological studies have demonstrated.

1.1.3. Genetics

Family studies have suggested that there are hereditary factors underlying OCD[48]. First-degree relatives of patients with OCD are 3 to 12 times more likely to develop OCD[49]. The risk of OCD among relatives of OCD probands increases proportionally to the degree of genetic relatedness[50]. Twins studies in families with OCD have a long history[51] – the first one was performed in 1929 – and overall have estimated the heritability of obsessive-compulsive symptoms within a range of 45% to 65%[52]. Monozygotic twins have the highest concordance rates – between 80 and 87% - followed by dizygotic twins, with concordance rates between 47 and 50%[52]. However, the sample sizes from twin studies are usually too small to allow for accurate heritability estimates [53].

Numerous groups have tried to localize the contribution of specific genes to the development of OCD. The most studied genes are associated with the glutamate system (glutamate transporter SLC1A1 or GRIN2B, encoding an NMDA-type glutamate receptor subunit), the serotonergic system (serotonin transporter SERT, serotonin receptors 5HT1B and 5HT2A) and the dopaminergic system (dopamine transporter DAT, dopamine receptors DR1, DR2, DR3, DR4)[54,55]. Association between OCD and immunity-related genes has also been found: a specific allele of the myelin-oligodendrocyte glycoprotein precursor (MOG) gene seems to increase OCD risk[56], while a specific TNF gene polymorphism seems to be protective[57]. The SLC1A1 gene is the only candidate gene that has been found in multiple independent samples, although the specific-associated polymorphism has varied[58].

Two OCD genome-wide association studies (GWAS) have been published by independent OCD consortia, the International Obsessive-Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and the OCD Collaborative Genetics Association Study (OCGAS)[58,59]. In the first of these studies, no single-nucleotide polymorphisms (SNP's) were found to be associated with OCD at a genome-wide significance level[58]. In the second GWAS, the smallest P-value was detected for a SNP on chromosome 9p23, in close proximity to the protein tyrosine phosphate receptor D (PTPRD) gene, a member of the receptor protein tyrosine phosphatase family that regulates transmembrane signaling molecules and that promotes glutamate receptor differentiation pre-synaptically[59]. In 2018, a meta-analysis of the two consortia (including a total of 2688 OCD patients and 7037 genomically-matched controls) has found an association with several glutamatergic system genes (GRID2 and DLGAP1) to add to other consistently implicated genes affecting this neurotransmitter system (SLCL1A1 and GRIN2B49, as mentioned previously)[53]. As in other psychiatric disorders, genetic studies have contributed to our understanding of the neurotransmitters which may be involved, but no specific genes have been found to cause the disorder and they have not led to new treatment approaches.

1.1.4. Neuroimaging and neuromodulation

Research using functional neuroimaging in OCD has shown a high degree of concordance among different studies that is probably the highest among all psychiatric disorders[60]. Regarding structural imaging, studies using a region-of-interest approach show a decreased volume of the orbitofrontal cortex (OFC) and of the anterior cingulate cortex (ACC)[61]. Whole brain-based analysis using voxel-level analysis methods (voxel-based morphometry, VBM) have confirmed the reduced OFC and ACC volume and revealed increased striatal volume, as well as decreased volume of the parietal cortex[62–64]. Very recently, the ENIGMA project, a consortium meta- and mega-analysis, published its first results. In this study, which compiled the largest number of OCD brain scans ever analyzed, adult OCD patients show a larger volume of the pallidum[65]. OCD patients also show smaller hippocampal volume, but that seemed to be driven by comorbid depression[65]. The cortical branch of the ENIGMA study, which used surface-based analysis instead of voxel-based morphometry, has found lower surface area in the transverse temporal cortex and a thinner inferior parietal cortex[66]. Functionally, the most consistent abnormality in OCD, both in positron-emission tomography (PET) and functional magnetic resonance imaging (fMRI) is increased activity in the medial and lateral orbitofrontal cortex, both in children and in adults, at rest or during neutral states[61,67–70]. There is also strong evidence pointing towards dysfunction of the caudate nucleus, particularly bilateral hyperactivity of the caudate head, again both in children and in adults (also at rest)[61,67–69,71]. Some studies also find hyperactivity in the ACC, both in rest and in symptom provocation[70,72,73]. The findings from functional neuroimaging studies have led to the proposal to the cortico-striato-thalamo-cortical (CSTC) dysfunction model of OCD at the turn of the century (Fig. 1)[74,75].

This model is based on what is known about the basal ganglia. The basal ganglia consist of four main strutctures: striatum, globus pallidus, substantia nigra and subthalamic nucleus[76]. The striatum is separated by the internal capsule into the caudate (dorsomedial striatum in rodents) and the putamen (dorsolateral striatum in rodents). The globus pallidus has two functionally different segments (internal and external), as well as the substantia nigra (*pars reticulata* and *pars compacta*). The organizing principle of the cortico-basal ganglia circuits is that they are a set of parallel, partly segregated, multi-synaptic circuits that begin with a cortico-striatal glutamatergic projection from the cerebral cortex to the dorsal striatum (Fig. 1, top panel)[74,75,77]. The striatum then makes a GABAergic inhibitory projection to the internal segment of the globus pallidus and to the substantia nigra *pars reticulata*. The internal segment of the globus pallidus is the major output structure of the basal ganglia, projecting mainly to the thalamus, from where recurrent projections head to the original cortical area where each specific loop originated[74,75,77]. This is called the direct pathway, in opposition to an indirect pathway, which originates in the dorsal striatum but has a GABAergic inhibitory projection to the external segment of the globus pallidus. In the indirect pathway, the external segment of the globus pallidus sends another inhibitory projection to the subthalamic nucleus, which then projects to the globus pallidus / substantia nigra with an excitatory glutamatergic synapse. In healthy persons, the direct pathway facilitates thalamic dishinibition and, consequently, stimulation of the original cortical area, in a feed-forward circuit that facilitates movement or whichever neural process instated in the cortical area from which it receives projections. The indirect pathway has an inhibitory function by modulating the activity of the direct pathway. Different areas of the cerebral cortex project in a highly topographic way onto the striatum and the topographic termination pattern establishes functional domains that are replicated throughout the basal ganglia–thalamocortical circuits through highly topographic projections at each synaptic relay[76,77]. The different loops that pass through the basal ganglia are named after the presumed functions of the regions of the frontal cortex from which they originate: the skeletomotor loop (originating in the primary motor cortex, premotor cortex and supplementary motor area), the oculomotor loop (originating in the frontal eye fields and in the supplementary eye fields), the prefrontal/executive/associative loop (originating in the dorsolateral prefrontal cortex and the lateral OFC) and the limbic/emotion loop (originating in the medial OFC and the anterior cingulate cortex). According to the CSTC dysfunction model of OCD, there is an excessive activity of the direct pathway and a diminished activity of the indirect pathway in loops starting in the OFC and in the ACC[74,75].

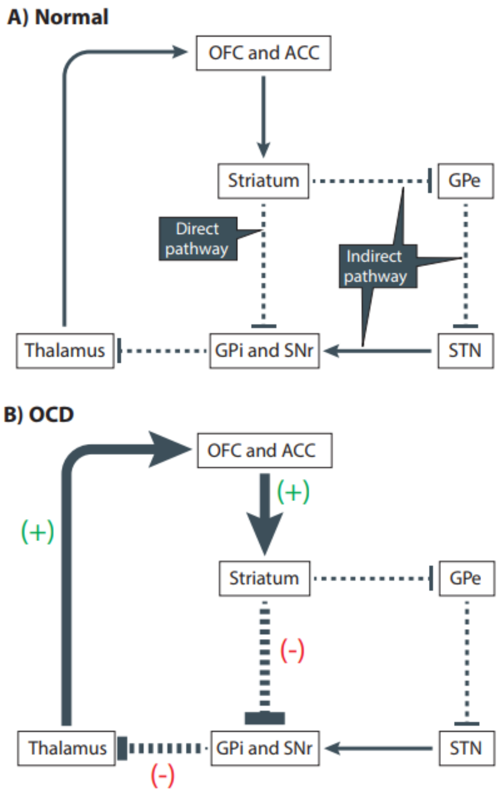


Figure 1. Cortico-striato-thalamo-cortical (CSTC) dysfunction model of OCD. A) In the normally functioning CSTC loops which start in the OFC and the ACC (limbic/emotion loops), glutamatergic inputs from these areas excite the striatum. Striatal activation generates

GABAergic inhibitory signals to the internal part of the globus pallidus (GPi) and to the substantia nigra pars reticulata (SNr) through the direct pathway. This will decrease the inhibitory GABA signals from the GPi and the SNr to the thalamus and will thus result in excitatory glutamatergic signals from the thalamus to the OFC and ACC. In the indirect pathway, the striatum inhibits the external part of the globus pallidus (GPe), which will decrease its inhibition of the nearby subthalamic nucleus (STN). The STN will then be released to excite the GPi and the SNr, causing thalamic output inhibition. **B)** According to the CSTC dysfunction model, OCD patients have an unbalance between the direct and indirect pathway in the CSTC loops which start in the OFC and the ACC, with excessive activity in the direct pathway and decreased activity in the indirect pathway. Adapted from Pauls et al., 2014[55], Saxena & Rauch, 2000[74] and Milad & Rauch, 2012[75].

Numerous studies have shown a reduction in metabolic activity in the OFC, caudate and ventrolateral prefrontal cortex post-treatment relative to pre-treatment in clinical trials assessing the effects of pharmacotherapy[78–81] or CBT[79,82,83] in OCD patients, which provides support for the CSTC model. Neuromodulatory treatment approaches which target the CSTC circuits have been tried in the last decades with promising results. Deep brain stimulation (DBS), a neurosurgical procedure in which electrodes are implanted deeply in the brain to convey direct electric current to specific brain regions has shown some efficacy in treatment-refractory OCD[84,85]. The most commonly targeted areas are the ventral striatum / nucleus accumbens[86,87], the anterior limb of the internal capsule[88,89] and the subthalamic nucleus[90,91]. Transcranial magnetic stimulation (TMS) has also shown promising results, with the advantage of being non-invasive. A recent meta-analysis has concluded that the target areas with strongest evidence of a significant effect are the dorsolateral prefrontal cortex (DLPFC) and the pre-supplementary motor area[92]. A less restrict meta-analysis added the OFC to this areas[93]. The United States (US) Food and Drug Administration (FDA) approved DBS as an humanitarian device exemption for treatment-refractory OCD in 2008 and cleared a variant of TMS (deep TMS)[94,95] that stimulates a broad prefrontal area in 2018.

1.1.5. Neuropsychology

Neuropsychology has contributed amply to our understanding of obsessive-compulsive symptoms and has shown that OCD patients display an impaired performance in several tasks and paradigms. The deficits most consistently associated with OCD are cognitive inflexibility – (both in reversal learning and in attentional set-shifting tasks) and motor impulsivity (reflected in motor prepotent motor disinhibition in stop-signal reaction time task [SSRT])[96]. However, reversal learning has appeared to be intact in one study[97] and motor inhibition in another[98]. Also, deficits in other dimensions of executive function have also been reported. Deficits in the Tower of London task, which measures planning capacity, have been reported and are well replicated[99,100]. Gambling tasks such as the Iowa Gambling Task show inconsistent results: a number of studies have found deficits in the Iowa Gambling Task[99] , yet another study has found no deficit in the same task[101]. Curiously, in a variant called the Cambridge Gambling Task, which does not require associative feedback learning, OCD patients showed no deficit[102].

Several researchers have tried to define neuropsychological endophenotypes associated with OCD. An endophenotype is a measurable component unseen by the unaided eye that lies halfway in the causal pathway between the clinical disorder and the underlying genotype[103,104]. It is a state-independent marker that does not include any of the symptoms that are necessary for the diagnosis of a particular disorder[104]. A valid endophenotype of a disorder should be present in unaffected family members of index cases. Endophenotypes represent simpler clues to genetic underpinnings than the disease syndrome itself, promoting the view that psychiatric diagnoses can be decomposed or deconstructed, which can result in more successful genetic analysis and a better clinical classification[104]. Compared with control subjects, OCD patients and their unaffected family members have worse performance on attentional set-shifting[105,106], motor inhibition[105,106], error-monitoring[107,108], planning (as assessed by the Tower of London task)[99,109] and delayed (verbal and non-verbal) memory tests[106,110], qualifying these neuropsychological measures as potential endophenotypes for OCD. Unfortunately, for many of these measures, negative studies also exist, suggesting that their sensitivity (and specificity) may be low[68,97,98].

A number of studies have tried to combine neuropsychological with neuroimaging approaches. Behavioral impairment on motor inhibition tasks such as the Stop-signal reaction time task (SSRT), occurring predominantly in OCD patients and unaffected relatives, has been significantly associated with reduced grey matter volume in the OFC and right inferior frontal regions, and increased grey matter volume in cingulate, parietal and striatal regions[111]. Combining functional MRI and neuropsychology has also proved to be a valuable strategy. For instance, in a seminal work by Chamberlain and colleagues these authors used a cognitive flexibility paradigm adapted for fMRI and were able to show that patients with OCD and their unaffected first-degree relatives exhibit under-activation of the bilateral lateral OFC during reversal of responses[112].

The main limitation of neuropsychological approaches to OCD is the lack of specificity of most findings, even those that have resulted from combining neuropsychological tasks and fMRI. Many of the most replicated impairments have also been described in other psychiatric disorders, from attention-deficit hyperactivity disorder to schizophrenia. As a consequence, while these approaches have unquestionably contributed to a much more profound understanding of the neurophysiology of OCD, we still lack a reliable biomarker for the disorder that proves able to discriminate OCD patients from patients with other neuropsychiatric disorders, which is sensitive enough to therapeutic interventions to be used as a reliable marker of clinical improvement , and which can be mapped onto brain circuits or function.

In the last decade, several authors focused on the fact that the CSTC loops that seem to be altered in OCD are well known by their role in supporting how the brain learns from action. One of the most distinctive features of OCD is the powerful urge felt by patients to perform specific acts, despite having full insight into how senseless and excessive these behaviors are, and having no real desire for the outcome of these actions. We will now review what is known about action learning in the brain and how this seems to be dysfunctional in OCD.

1.2. Action learning in OCD

1.2.1. Instrumental conditioning

The first studies about action learning in OCD patients[113] used tasks which were inspired by the animal literature of *instrumental conditioning*. The roots of instrumental conditioning date to the experiments performed by Edward Thorndike in the transition from the 19th to the 20th century[114,115]. These experiments, mainly performed in cats (but also in dogs, chicken and monkeys), consisted of putting the animal in what were called "puzzle boxes" with a closed door which would open only after the animal performed a specific sequence of actions (e.g. depressing a platform, then pulling a string, and then pushing a bar up and down). Thorndike observed that the time that each animal took to open the door diminished with successive experiments and formulated what came to be known as the *Law of effect*. This principle states that actions which are followed by pleasurable consequence become most likely to be repeated and actions which are followed by negative consequences become less likely to be repeated[114,115]. This law described what is generally known as trial-and-error learning and represents an enduring principle of learning. It had a major influence on Clark Hull[116] and the behaviorists, for whom instrumental learning resulted from stimulus-response bonds which could be strengthened or weakened by subsequent reinforcement (here meaning "anything that modifies the probability of an action"). B. F. Skinner, one of the most prominent behaviorists, did not agree with the associative linkages proposed by Hull, but focused on the selection of spontaneously emitted actions[117–119]. He introduced the term *operant conditioning* to emphasize the role of actions on the environment and had a very important contribution by developing what came to be known as the Skinner Box, in which the animals could press a lever to obtain rewards which were delivered according to specific rules called reinforcement schedules[117–119].

Stimulus-response theories, which were the core of experimental psychology along the first half of the 20th, argued that behavior reflected the development of an associative structure according to which a representation of the stimulus context became, with increasing experience, more strongly connected to a motor system which would generate behavioral responses. Some authors, however, were fiercely opposed to these ideas. The most well-known was Edward Tolman, who wrote that S-

R learning resulted in an animal coming to respond more and more ‘‘helplessly’’ to a succession of stimuli that ‘‘call out the walkings, runnings, turnings, retracing, smellings, rearings and the like which appear’’[120]. Tolman argued strongly that S-R was very poor to explain animal behavior and proposed that animals learn a maze task by forming a *cognitive map* of the environment, which then provides the necessary guidance mechanism for the observed learning. Tolman’s view was based on experiments performed by Blodgett[121], examining the nature of learning that occurs in the absence of the driving force of reinforcement. Blodgett found that an animal exploring a maze environment without experiencing a reinforcing reward contingency, can still be shown to be engaging in what is known as latent learning[121]. Latent learning is ‘‘unmasked’’ when the animal is subsequently tasked to navigate toward a rewarded goal state in this same environment. Animals who are previously exposed to the maze show facilitation in learning relative to naive animals, suggesting that the preceding nonrewarded exposure epochs foster the development of a cognitive map that aids future attainment of the rewarded goal location[122].

The debate between “pure behaviorists” and “cognitivists” only became partially softened in the 80’s, with the experiments of Dickinson, Adams and Rescorla[123–127]. These authors cleverly manipulated the relationship between actions and its consequences in order to distinguish stimulus-response (which they named *habitual*) from action-outcome (which they named *goal-directed*) control of behavior (Fig. 2). A specific behavior is considered goal-directed if it reflects knowledge of an association between an action and its outcome and takes into account the motivational value of the action[128,129]. In previously conditioned animals, this may be tested by two procedures: outcome devaluation and contingency degradation (Fig. 2)[123,125]. *Outcome devaluation* consists in manipulating the value of the reward, e. g. by pairing the reward consumption with lithium chloride-induced illness or by allowing free access to the reward outside the conditioning context. *Contingency degradation* consists in manipulating the contingency between the performance of an action and the delivery of the reward, e. g. after a period of training in which there is a specific contingency between pressing a lever and obtaining a reward (for example every 10 presses, on average), suddenly rewards start being delivered only when animals refrain from pressing. *Goal-directed behavior* is sensitive to both of these manipulations. On the other hand, an action can be rendered habitual through repetition[123,124,130]. In *habitual behavior*, a specific stimulus comes to elicit automatic responses which are independent of the consequences of the action (autonomous from the outcome) (Fig. 3). This makes habitual behavior insensitive to manipulations of the action-outcome contingency or of the value of the outcome[125]. Habitual behavior is sometimes said to be controlled by its antecedent stimuli while goal-directed behavior is controlled by its consequences[125,131]. Goal-directed control has the advantage that it can quickly change the animal’s behavior when the environment changes its way of responding to the animal’s actions[132]. Habitual behavior has the advantage of responding fast when the animal is accustomed to the environment, but it is unable to quickly adjust to changes in the environment[132].

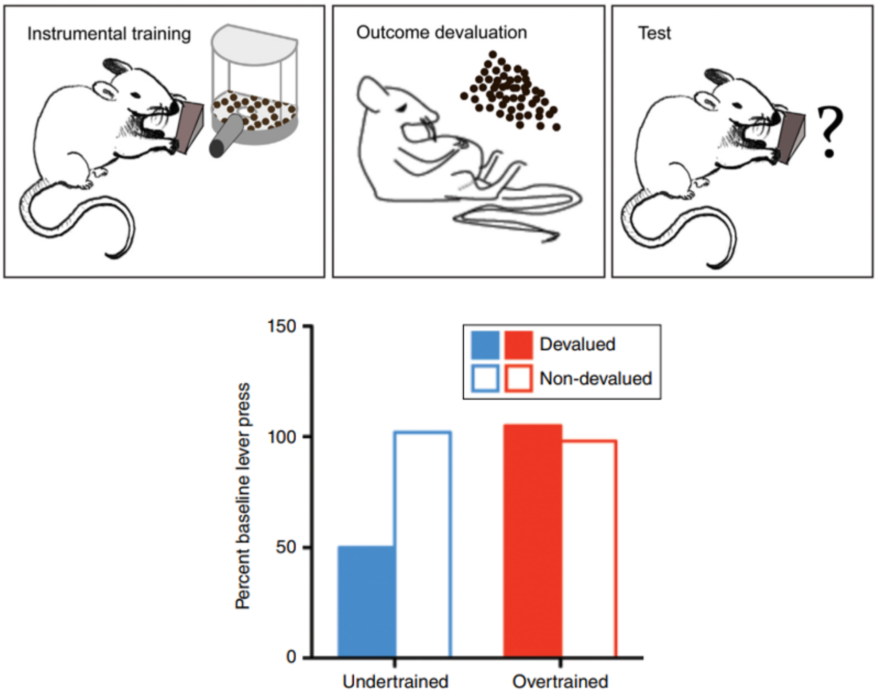


Figure 2. Distinguishing goal-directed from habitual behavior using outcome devaluation in rodents.

A) Left: hungry rats are trained to press a lever to obtain food. Middle: Feeding the animal to satiety devalues the outcome of pressing the lever (which is food). Right: Rats are confronted again with the lever to test if they still press for the devalued outcome and compared with control rats for which the outcome has not been devalued. This last test is performed without food delivery (in *extinction*) to make sure that behavior is based on previously learned associations. **B)** With low amounts of training (Undertrained, blue), a devaluation effect can be seen: animals reduce their responding for a devalued outcome (e. g., a food that had previously been fed to satiety), but continue to respond if the outcome is still valuable (i.e., has not been fed to satiety; Nondevalued). However, after extended instrumental training (Overtrained, red), they are insensitive to outcome value and continue to respond for a devalued outcome, reflecting a transition from goal-directed to habitual behavior. Adapted from Adams, 1982; Daw & O'Doherty, 2013 and Lingawi et al, 2015[124,133,134].

We know from our own personal experience that an action which starts out as goal-directed can turn into a habit with enough repetition (Fig. 4). Instrumental conditioning paradigms have shown that this also happens in rodents[124]. In a now classical experiment, Adams trained two groups of rats to press a lever for food rewards – crucially, one group was trained until they made 100 rewarded presses and the other group was trained until they made 500 rewarded presses (Fig. 2). After such training, an outcome devaluation procedure was carried out prior to a test (carried in extinction, to ensure behavior was based on previously learned associations) to analyze if the rate of pressing was reduced. Devaluation strongly reduced lever-pressing in the low-training group but the rate of lever-pressing in the extensive-training group was maintained, suggesting that the rats with minimal training remained goal-directed while the overtrained rats had developed a habit.

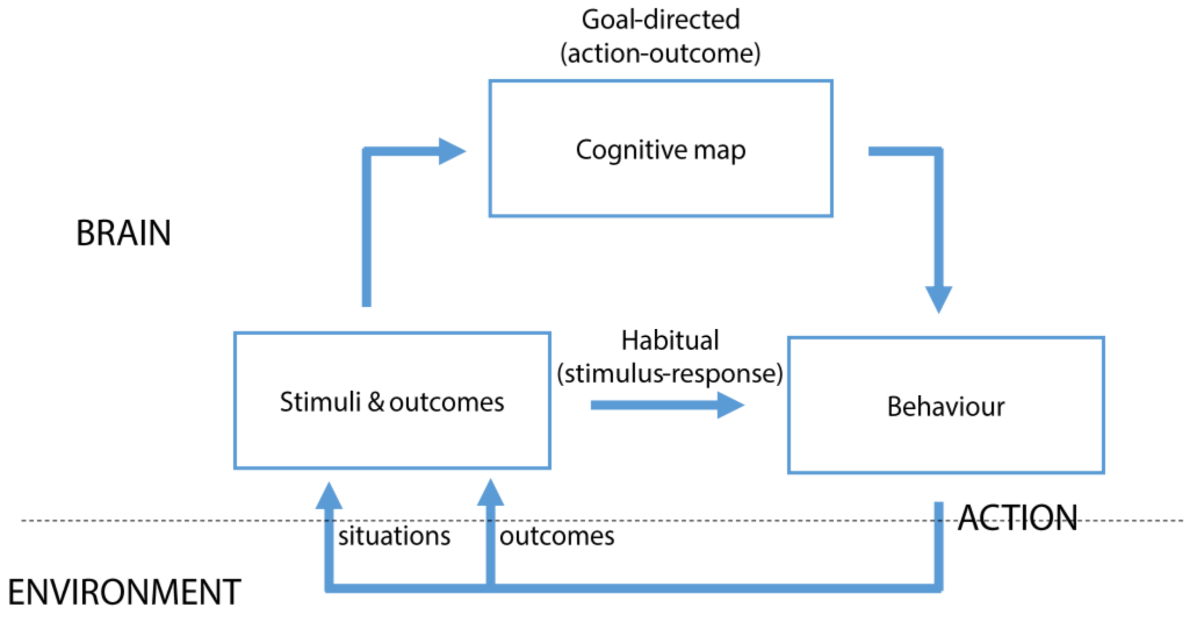


Figure 3. Distinct action control systems identified in instrumental conditioning experiments.

Instrumental conditioning paradigms, typically performed in rodents, have suggested that the brain can select actions using two different systems – goal-directed and habitual. In goal-directed behavior, animals establish an association between an action and its outcome. In habitual behavior, actions are stamped in by reinforcement of associations between stimulus and action (or response).

However, it has also become apparent that it is not only the extent, but also the nature of training that underpins habit learning. In fact, Adams and Dickinson also examined the pattern according to which the reinforcer was provided during training[123–125,130]. They discovered that schedules in which rewards were delivered on the first action after a specific interval — either a fixed interval or a variable interval schedule – were more prone to lead to habitual behavior than schedules where the absolute number of responses is the deciding factor for obtaining reward (fixed or variable ratio schedules). This is a commonly used experimental manipulation for generating goal-directed or habitual behavior[135], although it is not totally clear which of the factors that distinguish the schedules is responsible for this effect[129].

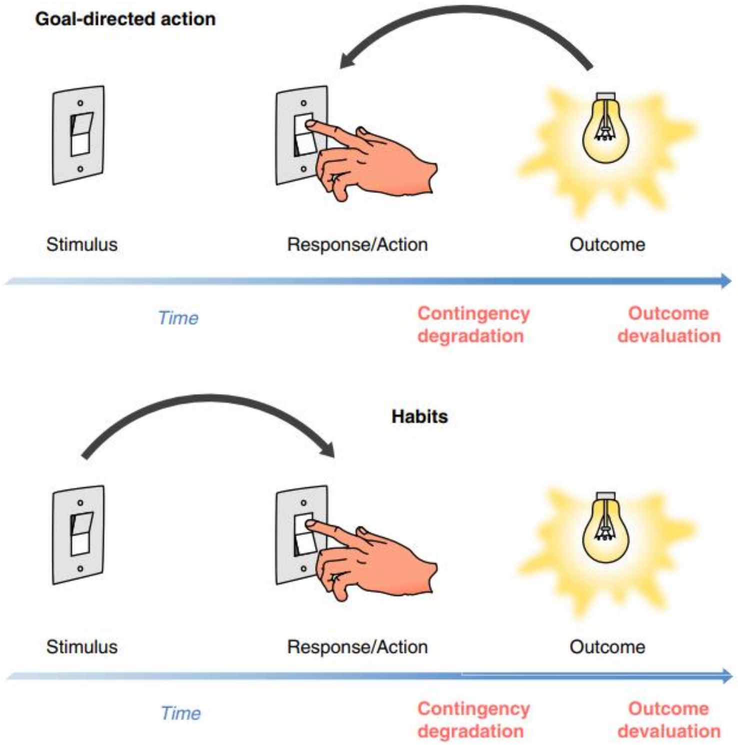


Figure 4. Goal-directed and habitual actions. Goal-directed actions are performed in order to obtain outcomes. So, in goal-directed behavior, animals learn that their actions have valuable consequences and have knowledge of instrumental contingencies of specific outcomes and their current motivational values. If the outcome (the light coming on) becomes less valuable (e. g., the room being well illuminated), an action controlled by a goal-directed system will be not be performed. Also, if the contingency between the action and the outcome becomes degraded (e. g., a movement sensor is installed and light comes on without need to press switches) the action will not be performed. In habitual behavior, the fundamental idea is that, through repetition and learning, environmental stimuli come to automatically elicit responses that were initially made spontaneously by the animal. So, habits are elicited by antecedent stimuli – in a stimulus-response fashion – and not performed to obtain future outcomes. Therefore, when someone develops a habit of pressing a switch, the action will be performed when looking at the switch, even if the outcome has lost its value or the relation between action and outcome has been modified. Adapted from Robbins & Costa, 2017[129].

It should be noted that instrumental (or operant) conditioning differs from pavlovian (or classical) conditioning[136] because in the former, delivery of a reinforcing stimulus is contingent on what the animal does, while in the latter, the reinforcing stimulus (i.e., the unconditioned stimulus) is delivered independently of the animal's behavior[128]. Thus, in pavlovian conditioning, the association that governs behavior is between a stimulus predicting an outcome (stimulus-outcome), which is different from the stimulus-response association that governs habitual behavior and the action-outcome association that governs goal-directed behavior[128]. Importantly, while outcome devaluation is used for instrumental actions to be established as goal-directed or habitual[125], pavlovian responses, both 'preparatory' (reflecting the motivational properties of the outcome) or 'consummatory' (reflecting the sensory properties of the outcome)[137] can also be sensitive to outcome devaluation [138]. Consequently, the assessment of goal-directed behavior should also include contingency degradation, which allows the distinction from the stimulus-outcome associations that governs pavlovian responses[128,139–141].

The neural circuits underlying each of the two systems involved in instrumental conditioning have been extensively studied in rodents. The structures which support goal-directed and habitual action strategies have been shown to differ, particularly at the level of the basal ganglia[142,143]. In rodents, as in humans, the striatum is also the entry station of the entire basal ganglia and serves as a hub for cortico-basal ganglia reentrant loops, being capable of integrating inputs from the cortex, the midbrain and other structures[144–146]. While the limbic loops that run through the ventral striatum (nucleus accumbens) seem to mediate pavlovian stimulus-outcome associations and responses, the loops that run through the dorsal striatum are more involved in the control of instrumental actions[141,142,147]. In the dorsal striatum the medial regium (DMS) extends ventrally until the nucleus accumbens and receives most of its input from associative cortical areas (like the caudate in humans) and the lateral regions (DLS) receive most of its input from sensorimotor cortical areas (like the putamen in humans)[148,149]. Lesion studies (or receptor blockade studies) combined with instrumental conditioning paradigms have shown that associative CSTC loops involving the DMS[150,151] and prelimbic subregion of the medial prefrontal cortex[152–154] are necessary for learning and performance of goal-directed behavior. At the same time, habit formation has been shown to depend on the DLS[155] and on the infralimbic subregion of the medial prefrontal cortex[156]. These parallel DMS/DLS corticostriatal circuits dynamically interact with each other[135,157], suggesting that competing CSTC circuits underlie the ability to switch between two different modes of performing the same action[158,159]. Interestingly, it has also been shown, in rodents, that the OFC has a crucial role in the shift between habitual and goal-directed behavior. Specifically, *in vivo* neuronal recordings reveal that OFC and DMS neurons become more engaged during goal-directed actions and DLS neurons become more engaged during habitual actions[135]. Also, chemogenetic inhibition of the OFC disrupts goal-directed actions, where optogenetic activation of the OFC increases goal-directed pressing for food rewards[135]. We should keep in mind that most of the neuroimaging differences between OCD and healthy subjects are precisely in the OFC and the caudate nucleus (which is the human equivalent of the DMS).

As the abovementioned experiments provided clear cut results about brain-behavior interaction in rodents, researchers tried to adapt these types of tasks to human subjects. The task which came to be known as the Fabulous Fruit Game was published in 2007 and is considered the first paradigm which tried to isolate action-outcome (goal-directed) and stimulus-response (habitual) contributions to human behavior[160]. In the training stage, participants were asked to respond to different pictured stimuli (fruits) in order to gain rewarding outcomes (points). In some trials (“congruent discrimination”) the stimulus was the same as the outcome in each component, whereas in other trials (“incongruent discrimination”) the stimulus of one component was the same as the outcome of the other component. In a subsequent (instructed) outcome devaluation test the authors assessed whether participants were able to flexibly adjust their behavior to instructed changes in the desirability of the outcomes. The authors show that incongruent trials renders actions resistant to outcome devaluation, suggesting habitual behavior, while after congruent trials actions are sensitive to outcome devaluation, suggesting goal-directed behavior[160]. Two years later the same group adapted this task for fMRI and showed increased activity in the ventromedial prefrontal cortex when behavior was sensitive to outcome devaluation (goal-directed)[161].

Valentin and colleagues developed the first instrumental learning task using devaluation through specific satiety in humans[162]. They scanned 19 healthy subjects in fMRI while they learned to choose actions which were associated with contingent delivery of liquid food rewards (orange juice, tomato juice or chocolate milk). After training, one of the foods was devalued by feeding subjects to satiety on that food. Afterwards, participants went back to the fMRI scanner and were exposed again to the same instrumental choice but in extinction (to make sure that behavior was based on previously learned associations and not on newly made associations). The authors found that the orbitofrontal cortex showed a strong modulation in its activity when comparing the devalued with the non-devalued action[162]. Another study used a contingency degradation procedure instead of outcome devaluation [163]. Here subjects were scanned while they pressed a button to gain small monetary rewards while the response-reward relationship changed over time. Blood-oxygen-level-dependent (BOLD) activity in the medial orbitofrontal cortex and in the caudate (the human analogous of the rodent DMS[128]) was higher in sessions when rewards were highly contingent on actions[163]. Also, the medial prefrontal cortex tracked local changes in action-outcome correlations[163].

Although these experiments shed some light on the human basis of goal-directed (action-outcome) behavior, they provided no information about the neural substract of habitual (stimulus-response) actions. To our knowledge, only Tricomi and colleagues tested habitual behavior in humans, using a design following the same principles of the animal literature[164]. In this experiment, healthy humans were given either a low (1 day) or a high (3 days) amount of training, with button presses leading to a rewarding outcome delivered in a variable-interval (VI) schedule of reinforcement. Performance was then assessed in extinction after outcome devaluation by specific satiety. The rewarding outcomes were food rewards and subjects were asked to fast before the experiment. Responding was totally self-paced, in contrast with the previously described human studies which were trial-based but in line with the animal paradigms. During the extinction test, participants in the 1-day group reduced their response rates during presentation of the cue linked to the devalued food, as would be expected if their behavior was goal-directed. On the other hand, participants in the 3-day group continued to respond for the devalued food, indicating that their behavior had become insensitive to changes in outcome value over the course of training. A within-group analysis of fMRI data from the extensively trained subjects comparing later sessions (when behavior was habitual) to earlier sessions (when it would likely have been goal directed) highlighted increased cue-related activity in right posterior putamen/globus pallidum, consistent with the rodent findings showing involvement of the dorsolateral striatum in habitual responding[164].

The conceptual proximity between habits, as described above, and the phenomenology of OCD – as well as the similarity between the brain areas involved in operant conditioning and in OCD – have motivated an interception between these research topics. As mentioned previously, OCD poses a paradox because patients recognize that their concerns (obsessions) are unrealistic and that their behavior (compulsions) is excessive or even absurd. However, their life can get stuck in repeating the same action – pressing a light switch or washing their hands – over and over again. Gillan and colleagues raised the hypothesis that goal-directed action control is compromised in OCD and that compulsive acts are driven by maladaptive habits[113]. To test this hypothesis, they used the Fabulous Fruit Game[160], which was previously shown to test the capacity to form action-outcome (goal-directed) associations. Importantly, the authors modified the design by adding a “slips of action” test, in which participants had to respond to stimuli that signaled devalued or still valued outcomes. In this “slips of action” test, the goal-directed and the habitual systems should compete for behavioral control. The authors also used a questionnaire to ask whether participants had developed explicit knowledge of the relationship between stimuli, actions and outcomes. They found that explicit knowledge of the relationship between actions and outcomes was impaired in OCD patients and that they made more errors in the “slips of action” test, suggesting a deficit in goal-directed control *and/or* an hyperactive habit system[113]. These results were interpreted at the time by the authors of the paper as a bias towards habitual behavior. However, as recently acknowledged by the same group[165], the slips of action test – similarly to other instrumental conditioning tests – does not allow to distinguish between low use of goal-directed behavior and high use of habitual behavior.

Among the different adaptations of instrumental conditioning tasks for use in humans, the one which is closer to the animal paradigms was the Tricomi task[164]. As that task has never been applied in OCD patients, in preliminary experiments I decided to replicate it in healthy subjects. In each training session, participants had access to two buttons, each one giving access to a food reward (“M&M’s” or “fritos”). Participants were asked to fast for at least 6h prior to the task. Each session consisted of consecutive task (20s-40s) and rest (20s) blocks, with each of the task blocks giving access to one of the two rewards. Task and rest blocks were pseudo-randomized, with an indication in each block of the active reward/button. Participants were told to press the buttons as much as they wanted and the availability of a reward was represented in the screen by the respective image. Rewards were probabilistically delivered according to one of two reinforcement schedules: one group was trained in a variable interval (VI) schedule of reinforcement and another group was trained in a variable ratio (VR) schedule of reinforcement. In the VI schedule, a reward was available, on average, every 10 seconds, being delivered on the first button press after the specific interval for that trial. In the VR schedule, a reward was available, on average, every 30 button presses. After the training phase, an outcome devaluation procedure was carried to test goal-directed vs. habitual behavior. Thus, free access was given to one of the food rewards to induce its devaluation by specific satiety. After this procedure, behavior was tested for a brief period in extinction (without rewards being delivered) allowing to test the effect of previous conditioning and not the effects of the reward *per se*. It would be expected that when conditioning induced goal-directed behavior, as demonstrated in rodents using VR schedules, more button presses would occur for the valued outcome that for the devalued outcome. Regarding habitual actions, which should result from VI schedules of reinforcement according to experiments in rodents, these differences would not be expected as the performance of actions would not be dependent on its consequences.

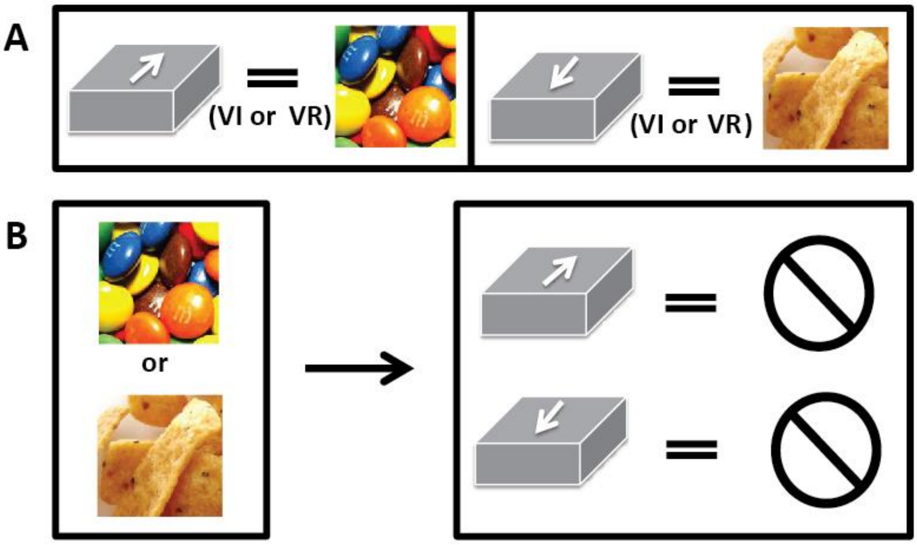


Figure 5. Adaptation of an instrumental conditioning experiment to human subjects.

A) In the training phase, subjects could press an arrow key to gain access to m&M's and another arrow key to get access to Frito's. Different stimuli on the screen signaled which food reward was available at each block. Rewards were delivered according to a 10-seconds Variable Interval or according to a 10-presses Variable Ratio reinforcement schedule. **B**, left panel) Outcome devaluation procedure through specific satiety: subjects could eat m&m's or Fritos until it was no longer pleasant to them. B, right panel) Extinction test: the discriminative stimuli were again shown in the screen and the participants could press the keys. During this 3-minute (extinction) test, responding no longer resulted in rewards. (Castro-Rodrigues et al., unpublished).

The operant conditioning task was applied with a VI schedule in 24 healthy volunteers, from which 14 completed a short-training protocol (two 8-minute sessions in one day) and the remaining 10 completed a long-training protocol (three 8-minute sessions in three consecutive days). After devaluation of one of the outcomes, there was a decrease in hunger, and a decrease in pleasantness of the devalued outcome, when comparing pre-devaluation with post-devaluation (Fig. 5). Pleasantness of the non-devalued outcome did not change. An outcome devaluation index was calculated as a proportion between the number of actions performed to obtain the devalued outcome and the number of actions performed to obtain the non-devalued outcome. In agreement with previous findings [164], the devaluation had a behavioral impact in the short-training group (devaluation index < 1) but not in the long-training group (devaluation index ≈ 1). In short, VI schedule of reinforcement lead initially to the acquisition of goal-directed behavior that with extended training gives rise to habitual behavior. However, this transition was unstable as it was not verified in the second half of subjects who were tested. In another group of 17 subjects, a VR schedule was used – in 8 of them with short-training and 9 with long-training. The devaluation effects in hunger and pleasantness scales were similar to the ones described above. However, and in disagreement with the findings in animal models, the devaluation index did not demonstrate the acquisition of goal-directed behavior, independently of the duration of training.

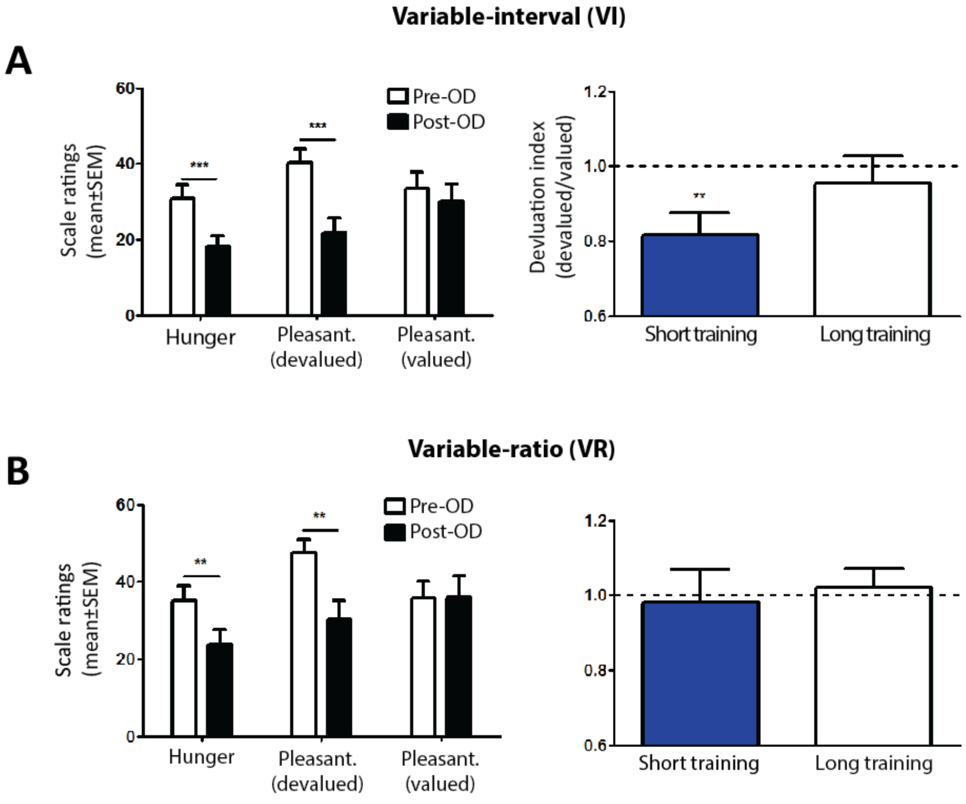


Figure 6. Results from the adaptation of an instrumental conditioning paradigm to healthy human subjects. A, left panel) Self-report rating scales for the group trained under the VI schedule of reinforcement. A, right panel) Devaluation index (proportion between key presses for devalued outcome and key presses for non-devalued outcome in the extinction test) in the VI low-training group and in the VI high-training group. B, left panel) Self-report rating scales for the group trained under the VR schedule of reinforcement. B, right panel)

Devaluation index in the VR low-training group and in the VI high-training group. VI = Variable-Interval. (Castro-Rodrigues et al., unpublished).

De Wit and colleagues in Amsterdam, and the Robbins group in Cambridge reported last year that they were incapable of replicating the overtraining findings from Tricomi and colleagues[165]. They have shown in five separate experiments, with three different learning procedures (including the Fabulous Fruit Game and an avoidance paradigm) that extended training does not significantly enhance habits in humans[165]. These five experiments include two failed independent replications of the report by Tricomi that overtraining induces habitual behavior in humans[165]. Interestingy, the only difference between these replications (and the replication I also tried to perform) is that the original study was performed inside an fMRI scanner. This has led others [129] to propose the habitual behavior shown there could have been induced by the stress of being in a confined space, as it has been shown that stress shifts goal-directed to habitual behavior in rodents [166]. The authors of this paper also point out that, in the analyes performed by Tricomi and colleagues, it is not totally clear that the BOLD signal in the putamen is directly related with behavioral sensitivity to outcome devaluation[165]. It thus seems that it is not trivial to experimentally induce habits in healthy humans as a function of behavioral repetition and that there currently exists no procedure that can reliably be used to do so[165].

Moreover, outcome devaluation has clear limitations as a paradigm for experimental human neuroscience. Firstly, the critical devaluation test during which behavioral strategies are dissociated must be short, because it is performed in extinction, limiting the number of choices or actions performed. Secondly, devaluation is a unidirectional single-opportunity manipulation of value. In fact, across the last decade, some authors have concluded that operant conditioning tasks, although very useful to study food rewards in rodents, work less well for other types of rewards and in humans[129,132]. These types of paradigms require extensive amounts of training before there is a possibility of testing for use of goal-directed vs. habitual behavior. Furthermore, they do not allow for separate assessment of habitual and goal-directed processes – meaning that either failures of goal-directed control or excessive habit formation could drive the failures to adjust action performance after devaluation. Finally, these experiments generate relatively small size datasets, with few trials per individual for analyses[132]. There is thus need for behavioral paradigms that generate large action selection datasets, with parametric variation of decision variables. In recent years, multistep decision tasks, which are inspired by reinforcement learning theory, have been developed with the objective of better quantifying and studying how humans learn from actions.

1.2.2. Reinforcement learning

Reinforcement learning (RL) is a field of computer science and machine learning that studies how agents of any sort can learn by interacting with their environment[167]. At its origins in the artificial intelligence field, an agent was an artificial system, such as a robot or a computer program. More recently, however, it has been hypothesized that the brain may also be implementing behavior through RL algorithms[168–171]. In fact, although RL theory was developed by the community which studied artificial intelligence, it drew significant inspiration from psychological learning theory[167,171,172]. RL problems involve four important components: states, actions, transitions and rewards[132,167]. *States* are contexts or stimuli in the environment that the agent observes and which are the basis for making choices; *actions* are behavioral choices made by the agent, that may or may not be available at each state; *transitions* are modifications from one state of the environment to another which are occasioned by actions; *rewards* are the basis for evaluating choices and can be food, water, money or any scalar measure of performance of the agent[132,167]. In the typical RL framework, a learning agent observes the state of its environment repeatedly and then chooses an action to perform. This action will change the state of the world (according to a transition function, which is typically unknown to the agent) and will probabilistically lead to a payoff (a scalar reward signal). *Utilities* quantify the subjective immediate worth of states in terms of rewards and punishments and depend on the motivation of each subject (for example, food has a higher utility when hungry)[132]. Two important definitions follow from that described previously: policy and value. In RL, *policy* is a mapping from states to actions (which is typically probabilistic) and *value* is the expected long-term sum of rewards[167]. In the RL framework, the *goal* of the agent is to learn to choose the policy that maximizes the value function (leading to the highest long-run sum of rewards)[167].

Several methods to solve RL problems have been described[167], namely algorithms that specify how the agent's policy is changed as a result of its experience. A common way of classifying RL methods is by distinguishing model-based RL from model-free RL, where the term model refers to a mental, not a computational model. In fact, their main difference is whether or not these algorithms build a representation – a model – of the dynamics of the environment which can be used to simulate trajectories in a task (Fig. 7). A model-free system does not build such a model but relies on estimating the value of each state or action – it relies on stored values for state–action pairs (Fig. 7)[167]. These values are estimates of the highest return the agent can expect for each action taken from each state – generally speaking, they tell how "good" or "bad" it is to be in those states or to take specific actions. They are obtained over previous experience in the environment. When the action values have become good enough estimates of the optimal returns, the agent just needs to select at each state the action with the largest action value in order to make optimal decisions. This strategy is "model-free" because it has no representation of the environment's causal structure (i.e., the transition function between states and the reward function in each state). Instead, it incrementally builds a look-up table or function approximation from which values can be quickly computed. Typically, in order adjust to its expectations, a model-free system uses a *reward prediction error* as a learning signal. The most known example of a prediction error is the *temporal difference* (TD) predictor error[173]. In fact, model-free predictions are supposed to be of the long-run sequence of actions starting on one step, so the ideal prediction error would measure the difference between the total amount of reward that is delivered over the long-run and the amount of reward that is predicted. However, waiting to experience all those rewards in the long run is usually impossible. The TD prediction error obviates this requirement via the trick of using the prediction at the next step to substitute for the remaining rewards that are expected to arrive[167]. In any case, prediction errors are based on the rewards that are actually observed during learning and train predictions of the long-run worth of states, criticizing the choices of actions at those states accordingly. Further, the predictions are sometimes described as being *cached*, because they store previous experience.

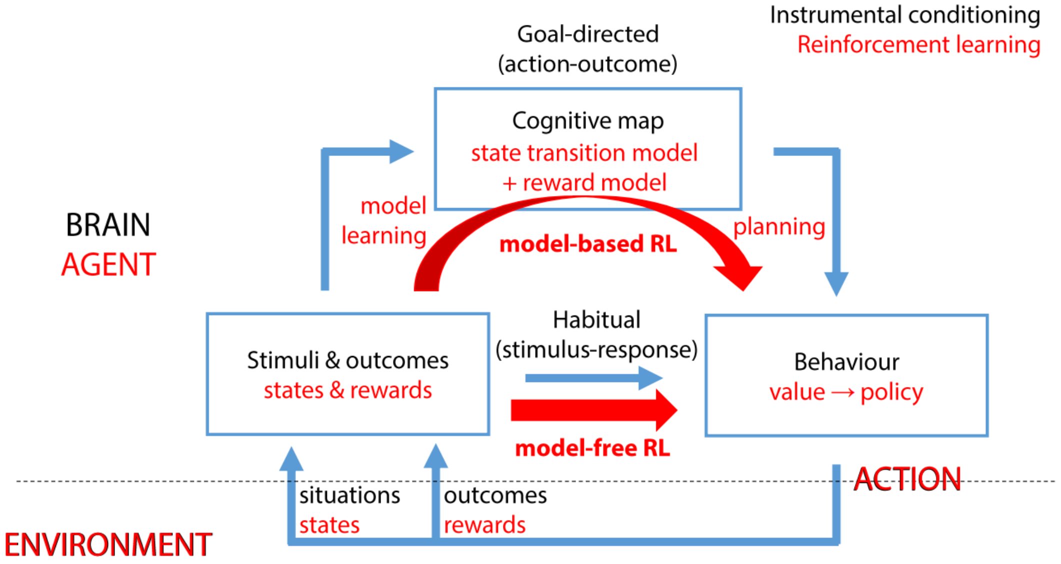


Figure 7. Parallels between instrumental conditioning and reinforcement learning.

In RL, situations are called states and outcomes are called rewards[167]. Policy is a mapping from states to actions ("which actions should be performed in each state"). There are two families of algorithms that can be implemented in order to choose the best policy: model-based and model-free[167]. Model-based RL works by building a model of the dynamics of the environment, just like the goal-directed action control system in instrumental conditioning. The model is then used to simulate possible trajectories and choose the action which has the highest value (expected long-sum of rewards). Model-free RL works by estimating the value of states and actions by trial-and-error and updates these estimates based on a reward prediction error (the difference between a reward that is being received and the reward that is predicted to be received). This ends up in repeating actions that lead to rewards in the past, just like habitual behavior in instrumental conditioning[158]. Due to its complementary strengths, it is advantageous for an agent to use both strategies: model-based RL in environments that change frequently and model-free RL in stable environments[132,158].

In model-based RL, the model is an internal representation of the environment which allows the agent to predict how the environment will change depending on the agent’s actions (Fig. 7)[132,167]. This environment model typically consists of a state-transition model and a reward model. The state-transition model is a decision tree which represents the probability of each state giving access to a different state depending on each possible action. The reward model associates the distinctive features of the goal boxes with the rewards to be found in each. A model-based agent can decide which action to choose at each state by using the internal model to simulate sequences of action choices to find a path yielding the highest return. In this case the return is the reward obtained from the outcome at the end of the path. After learning the model, a model-based agent can use it to construct a value function or policy – that process is called planning. Comparing the predicted returns of simulated paths is a simple form of planning (Fig. 8). When the environment of an agent using model-free RL changes, the agent has to experience the new characteristics of the environment in order to update its’ expectations[132,158,167]. For a model-free agent to change the action its current policy associates with specific a state, it has to move to that state, act from it and experience the consequences, probably several times. In contrast, a model-based agent can take into account information about environmental changes before having to experience their consequences.

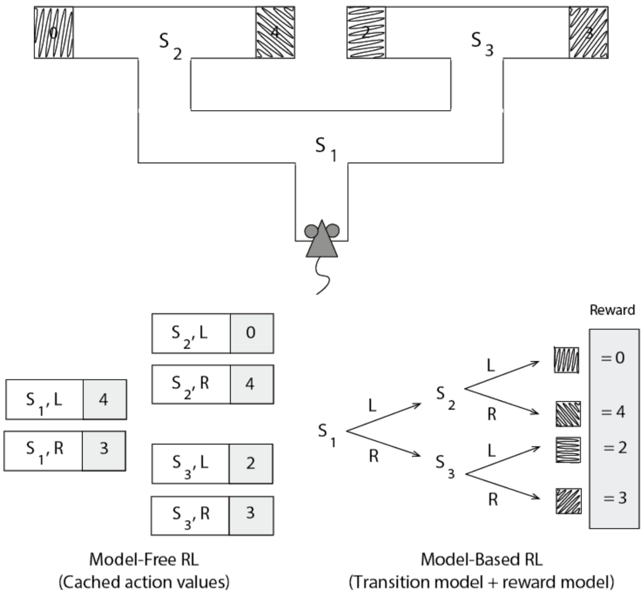


Figure 8. Model-based and model-free RL strategies to solve a sequential action-selection problem.

Top: a rat navigates a hypothetical maze with distinctive goal boxes at different end points, each associated with a reward having the value shown. Lower left: a model-free strategy relies on stored action values for all the state–action pairs obtained over many learning trials. To make decisions the rat just has to select at each state the action with the largest action value for that state. Lower right: in a model-based strategy, the rat learns an environment model, consisting of knowledge of state-action-state transitions and a reward model consisting of knowledge of the reward associated with each distinctive goal box. The rat can decide which way to turn at each state by using the model to simulate sequences of action choices to find a path yielding the highest amount of reward. Adapted from Sutton & Barto, 2018[171] and Niv et al., 2006[174].

In short, RL algorithms are a class of algorithms that have a narrow characterization: they try to maximize a specific cost function, the discounted sum of future expected reward[171]. Here discounted means that obtaining a reward now is better than obtaining a reward in the future. Such a function would clearly be important for animal survival and, as such, researchers hypothesized that such RL algorithms might be implemented in the brain. In the turn from the 20th to the 21st century, neuroscientists started to test the hypothesis that the brain is an evolutionary-shaped RL system on its own right, with studies in rodents, primates and humans[132,171]. These types of studies have sought for correspondence between signals in the brain, such as neuronal firing or BOLD activity, and signals which play fundamental roles in RL algorithms, namely terms in an RL equation or algorithm. In fact, one of the most transformative observations in this area of neuroscience was precisely the result of a study where neuronal activity was recorded in awake monkeys while receiving juice rewards in a pavlovian task[175]. It was observed that neurons in the ventral tegmental area (VTA) – one of the two main regions with dopamine-producing neurons, the other being the substantia nigra – began to signal the presence of a stimulus predicting reward, and stopped responding to the reward itself[175]. Also, if the predicted reward was omitted after learning, the same VTA neurons showed diminished responses at the time where reward was expected. These findings suggest that the (phasic) release of dopamine from the VTA to other areas (such as the OFC, the nucleus accumbens and the amygdala) may be signaling the presence of reward relative to its prediction, instead of simple reward delivery. This concept of a “prediction error” is precisely a term which is included in a model-free RL algorithm called the TD learning that I have mentioned previously. Initial human imaging studies that used RL methods to examine the representation of values and prediction errors mainly focused on model-free prediction and control, without exploring model-based effects[176–178]. These showed that the BOLD signal in regions of dorsal and ventral striatum correlated with a model-free temporal difference prediction error, the exact type of signal thought to be at the heart of reinforcement learning. Others have used special fMRI techniques to highlight the brainstem nuclei and, with designs mimicking the study with monkeys described above, have found the same prediction error in the VTA[179].

More recently, several experimental paradigms have provided as sharp a contrast between model-free and model-based for human studies, as animal paradigms have provided between goal-directed and habitual control. The first of these studies was performed by Gläscher and colleagues[180] and drove inspiration from

Tolman and his concept of cognitive maps[122]. The authors designed a two-stage Markov decision task that allowed to separate signals of a reward prediction error (RPE) during some trials from signals of a state prediction error (SPE) during other trials. A Markov-decision process (MDP) is a fundamental concept in reinforcement learning. MDP's are a classical formalization of sequential decision making, where actions influence not just immediate rewards, but also subsequent situations, or states, and through those, future rewards. Thus, MDPs involve the need to trade-off immediate and delayed reward, similarly to many real-world problems. As pointed out previously, the RPE is part of model-free algorithms and, in this study, the SPE was considered part of a model-based algorithm – by incorporating discrepancies between the learned model and observed state transitions. In order to dissociate SPE from RPE, volunteers were first exposed to the state space (their actions originated different states) without any rewards, providing an assessment of a pure SPE in a first fMRI session. Next, during a break, subjects were told reward contingencies and trained the reward mapping with a simple choice task. Then, they went back to the scanner and made choices to obtain rewards (here providing an assessment of the RPE). The authors found the presence of RPE's in the ventral striatum (which receives one of its major inputs from the VTA) and of SPE's in the lateral prefrontal cortex and in the intraparietal sulcus[180]. However, the contribution of model-free or model-based systems were tested separately and this design is closer to a latent learning paradigm than to a sequential action selection problem.

The original two-step task, first published by Nathaniel Daw and colleagues in 2011[181], was designed to encourage a balance between use of model-based and model-free RL. It is a two-stage Markov decision task in which, on around 200 consecutive trials, participants are required to make 2 choices to arrive to a rewarded or a non-rewarded outcome (Fig. 9). An initial choice between 2 options leads probabilistically to one of two 2nd-step "states". At the 2nd-step, another choice between two options is required, each of which is associated with a different chance of delivering a small monetary reward. Crucially, each 1st-step option leads more frequently to one of the 2nd-step states (a "common" transition), whereas it leads to the other state in a minority of the choices (a "rare" transition). Also, the chances of reward associated with the four 2nd-step options change slowly and independently across the trials. Because a model-based system is able to incorporate the probability of state-state transitions into its decision-making, while a model-free system is not, the predictions made by these systems are different after some combinations of events[181]. The model-free strategy is insensitive to the structure of the task and it will simply increase the likelihood of performing an action if it previously led to reward, regardless of whether this reward was obtained after a common or a rare transition. On the other hand, an agent using a model-based strategy would show differences in behavior following common and rare transitions. For example, when a model-free agent obtains a reward after a rare transition, it will choose the same first-step option on the next trial, since action values are updated based exclusively on the reward that follows the action[181]. In contrast, a model-based agent, who can represent the task transition structure, will be more likely to switch to the previously unchosen 1st-step option, since this behavior is more likely to lead to the 2nd-step state which was just rewarded. Using these predictions about first-step choice behavior, it is possible to infer the influence of the RL controllers in terms of the main effect of reward (model-free) and the interaction between reward and transition (model-based) on the probability of staying with the same 1st-step choice[181]. Using this task, Daw and colleagues have shown that humans use a mixture of model-based and model-free RL while performing the task (Fig. 9, panel D). The task has also revealed BOLD signals associated with model-based and model-free computations both in the ventral striatum and in the medial prefrontal cortex[181].

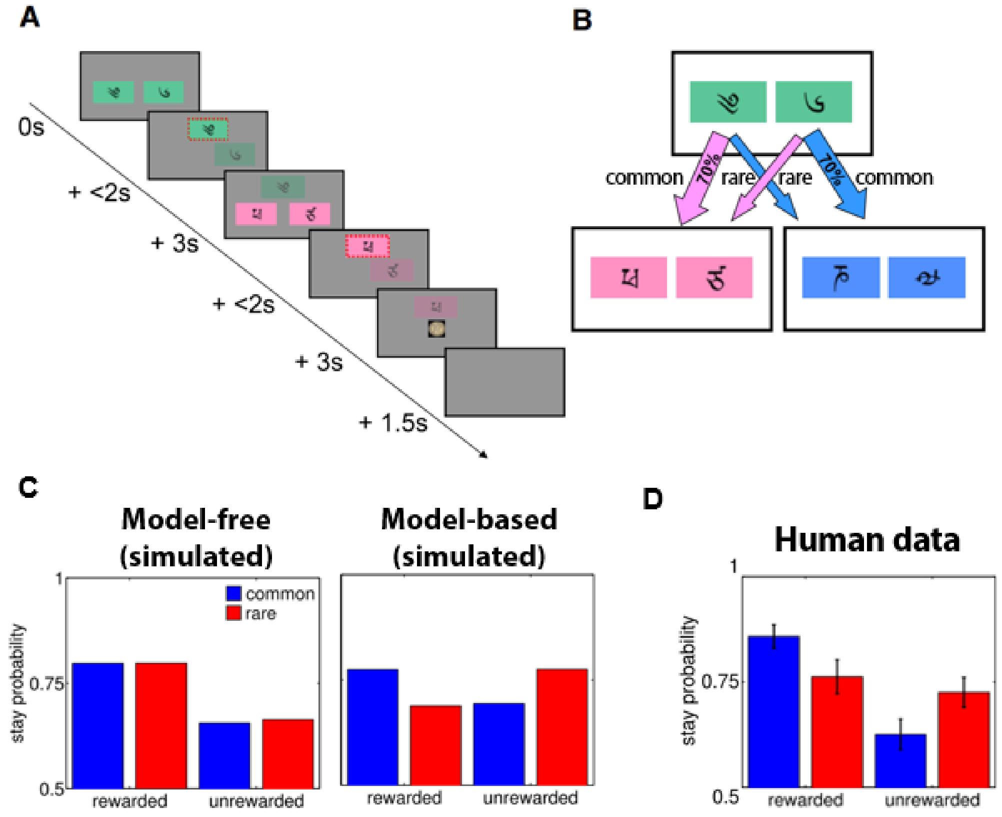


Figure 9. Original two-step task structure and results.

A) Trial events: On each trial, a first-step choice between two stimuli (identified by two tibetan characters) leads to a second-step choice which can give access to a small monetary reward (represented as a dollar coin). B) Task structure: Each first-step choice leads more frequently to one of the two second-step states (a "common" transition) and less frequently to the other (a "rare" transition"). To encourage learning, each of the four second-step possible options is associated with an independent reward probability which fluctuates slowly over trials. C) Probability of repeating the same first-step choice according to the events on the previous trial. Left panel: An agent which uses model-free RL will show higher probability of repeating the same first-step choice after trials that lead to a reward than after unrewarded trials, regardless of whether a common or rare transition occurred. Right panel: An agent which uses model-based RL will also take into account the transition that occurred, such that a rare transition will affect the value of the other first step choice and its stay-probability plot will show an interaction between the factors of reward and transition. D) Behavioral data from healthy subjects shows a hybrid model with hallmarks of both strategies. Adapted from Daw et al. 2011[181].

Wunderlich designed another sequential decision task, in order to explore the neural correlates of forward-based planning[182]. It was a minimax decision task which allowed to contrast forward planning with the correlates of extensively trained behavior[182]. It consists of a three-layer maze with the first-choice being made by the participant, the second-choice made by the computer and the third-choice again by the participant, then reaching a probabilistic reward. The results pointed towards a representation of forward planning values in the caudate (the human analogous of the rodent dorsomedial striatum[128]), a representation of extensively trained actions in the putamen (the human analogous of the rodent dorsolateral striatum[128]). Also, the ventromedial prefrontal cortex (vmPFC) was found to represent the value of the chosen option across both systems, suggesting that it may act to compare between them. Lee and colleagues, on the other hand, have tried to explore how control changes from a model-based (goal-directed) system to a model-free (habitual) system and vice versa[183]. They developed an arbitrator model which included three different levels of computation – learning; reliability estimation and reliability competition. In the first level, a model-free system generates a reward prediction error and a model-based system generates a state prediction error. At the second level, each of the systems calculates an estimate of the reliability of its specific prediction error. At the third level, these two reliability estimates in order to set a weight that regulates which of the two systems controls behavior. To test this model, the authors developed a two-stage Markov decision task in which participants gained access to colored tokens which gave access to monetary rewards. The task consisted of two types of trials – specific and flexible goal trials. Specific goal trials encouraged model-based RL while flexible goal trials favored the use of model-free RL. The state transition probabilities were also manipulated in order to favor model-based or model-free RL. Using this task, BOLD activity in the inferior lateral prefrontal and frontopolar cortex encoded the reliability signals and the output of a comparison between both signals[183].

Although they could seem similar, these types of tasks are different from classical tasks such as perceptual decision making or probabilistic reversal learning, where the only uncertainty about the outcome of each decision is whether reward will be directly delivered, making model-based prediction of future state and model-free prediction of future reward ineluctably confounded. However, neither the Glascher, the Wunderlich nor the Lee task were tested in clinical populations or subject to *in-vivo* neuromodulatory approaches (such as TMS). On the other hand, the two-step task became a popular approach to study the balance between use of model-free and model-based RL in healthy and in clinical populations. Importantly, Voon and colleagues found deficits in model-based control in OCD using the two-step task (Fig. 10)[184]. In the same study, patients with methamphetamine-addiction and with binge eating disorders also presented the same deficits[184]. Deficits in model-based control have also been found in other conditions such as schizophrenia and alcohol dependence[185,186]. Furthermore, Eppinger and colleagues studied age-related differences in model-free and model-based RL using the Daw two-step task[187]. Their results demonstrate age-related deficits in model-based decision-making, which were especially pronounced if a reward which was not expected indicated the need to shift the decision strategy – in this situation, younger adults explored the task structure while older adults showed some evidence of perseveration[187]. Also, in younger adults. high working memory (WM) capacity was associated with greater use of model-based RL and this effects was higher when the reward probabilities were more distinct[187]. Another group focused specifically in this effect of working memory and its relationship with the stress response [188]. These authors paired an acute stressor with the two-step task, assessed baseline WM capacity and used salivary cortisol to measure hypothalamic-pituitary-adrenal axis stress response[188]. They manipulated stress levels by using the cold pressor test task, an acute stress induction in which subjects submerged their arms in ice water for 3 minutes[188]. They found that the stress response attenuated the contribution of model-based but not of model-free RL. Furthermore, the stress-induced behavioral modifications were modulated by individual WM capacity – low WM-capacity subjects were more susceptible to detrimental stress effects than high WM-capacity subjects[188]. Following the hypothesis that model-based RL requires cognitive resources, the same author also demonstrated that having humans performing a secondary task leads to increased reliance on model-free RL strategies [189]. Moreover, it was also shown that, across trials, participants negotiated the trade-off between the two systems in a dynamic fashion, as a function of concurrent demands of executive function – and subject's latencies of choice reflected the computational expenses of the strategy they decided to use[189]. Others have used a similar approach to find that individual differences in processing speed covary with a shift from model-free to model-based control in the presence of above-average WM function[190].

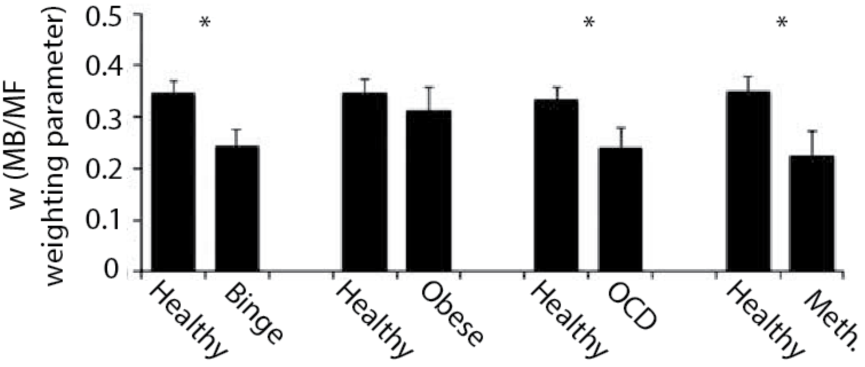


Figure 10. Performance of OCD patients, other clinical populations and healthy controls in the original two-step task. W is a weighting parameter in the computational model reflecting the balance between model-free (w = 0) and model-based control (w=1). Healthy = healthy volunteers, Binge = Binge eating disorder, OCD = obsessive-compulsive disorder, Meth = methamphetamine-dependent. Adapted from Voon et al., 2015 [184].

Other groups took a different approach and used non-invasive brain stimulation techniques to modify cortical activity and observe the effects of this manipulation on RL strategies[191]. These authors demonstrated that, when using the two-step task, it is possible to shift the balance between model-free and model-based control by disrupting activity in the dorsolateral prefrontal cortex (DLPFC) using theta burst transcranial magnetic stimulation[191]. They showed that disrupting activity in the right DLPFC leads to a dominance of model-free control, while disruption of left DLPFC impaired model-based performance only in those participants with low WM capacity[191]. The construct validity of the correspondence between goal-directed behavior and model-based RL and between habitual behavior and model-free RL has also been tested. Friedel et al.[192] used a devaluation paradigm[162] and the two-step task to address this question in healthy humans. There was a positive correlation between model-based control during the multistep task and goal-directed behavior in the outcome devaluation task. The authors concluded that a single framework may underlie these different operationalizations and that their findings support the construct validity of both approaches[192]. However, not all authors agree with the correspondence between habitual behavior and model-free RL[133].

In order to reach a better understanding of the processes underlying the use of reinforcement learning controllers in humans, both in healthy and clinical populations, sequential decision tasks need to be optimized. The two-step task, the paradigm that came to be used in several different neuropsychiatric disorders, presents some limitations. As in most human decision tasks, subjects performing the two-step task receive extensive prior instruction about task structure. Though there is extensive literature showing that instruction profoundly shapes human behavior in operant[193–195] and fear[196,197] conditioning, as well as value-based decision making[198–200], instruction effects have not been explored in multi-step tasks where model-based and model-free control can be dissociated. It therefore remains unclear how these different RL mechanisms contribute to action selection in situations where subjects must learn task structure directly from experience, and how providing explicit information about task structure modifies each system's computations. Also, the findings from the fMRI studies using sequential decision tasks reviewed above are not fully compatible. To address these questions, I developed a new sequential decision task and applied it in healthy subjects, in patients with OCD and, to control for the effects of medication and anxiety, in patients with mood and anxiety disorders (Chapter 3). I also adapted the same task for functional magnetic resonance imaging and designed a protocol to explore brain activity during performance of instructed and uninstructed sequential action choice (Chapter 4).

1.3. Aims

Aim 1 – Establish the criterion-related validity of the Y-BOCS-II for the diagnosis of OCD (Chapter 2)

Aim 2 – Develop a new sequential decision task to explore instructed and uninstructed reinforcement learning strategies in OCD patients and controls (Chapter 3)

Aim 3 – Create a protocol to collect functional imaging data during performance of a new sequential decision task (Chapter 4)

Chapter 2. Criterion-validity of the Y-BOCS-II for diagnosis of OCD

2.1. Abstract

While the Yale-Brown Obsessive-Compulsive Scale Second Edition (Y-BOCS-II) is the gold-standard for measurement of obsessive-compulsive (OC) symptom severity, its factor structure is still a matter of debate and, most importantly, criterion validity for diagnosis of OC disorder (OCD) has not been tested. This study aimed to clarify factor structure and validity of the Y-BOCS-II.

We first validated and quantified the psychometric properties of a culturally adapted Portuguese translation of the Y-BOCS-II (PY-BOCS-II). The PY-BOCS-II and other psychometric instruments, including the OCD subscale of the Structured Clinical Interview for the DSM-IV, used to define OCD diagnosis, were administered to 187 participants (52 patients with OCD, 18 with other mood and anxiety disorders and 117 healthy subjects). In a subsample of 20 OCD patients and the 18 patients with other diagnoses Y-BOCS-II was applied by clinicians blinded to diagnosis.

PY-BOCS-II had excellent internal consistency (Cronbach’s α=0.96) and very good test-retest reliability (Pearson’s r=0.94). Exploratory factor analysis revealed a two-factor structure with loadings consistent with the Obsessions and Compulsions subscales. There was good convergent validity but divergent validity was acceptable at best. The area under the curve (AUC) of the receiver operating characteristics (ROC) curve suggested elevated accuracy in discriminating between patients with OCD and control subjects (AUC=0.96; 95% confidence interval [CI]: 0.92 - 0.99), that was retained in comparisons with age, gender and education matched controls (AUC=0.95; 95% CI: 0.91 - 0.99), as well as with patients with other mood and anxiety disorders (AUC=0.93; 95% CI: 0.84 - 1). Additionally, a cut-off score of 13 had optimal discriminatory ability for the diagnosis of OCD, with sensitivity ranging between 85 and 90%, and specificity between 94% and 97%, respectively when all samples or only the clinical samples were considered.

The PY-BOCS-II has excellent psychometric properties to assess the severity of obsessive-compulsive symptoms, reflecting obsessive and compulsive dimensions, compatible with currently defined subscales. Importantly, we found that a cut-off of 13 for the Y-BOCS-II, total score has good to excellent sensitivity and specificity for the diagnosis of OCD. However, we confirmed that the Y-BOCS-II has problems in divergent validity, particularly regarding symptoms of depression.

2.2. Introduction

Accurate assessment of OCD is critical due to its under-diagnosis, difficulty in establishing accurate diagnosis and need for careful and specific treatment planning and evaluation[201]. The Yale-Brown Obsessive-Compulsive Scale is a clinician-administered instrument, developed in 1989 to assess the presence and severity of obsessive-compulsive symptoms[26,202]. It is divided into a symptom checklist and a severity scale. The symptom checklist comprises 54 dichotomous items assessing current or prior presence of specific obsessions and compulsions. The severity scale consists of 10 items that quantify the impact of obsessions and compulsions identified using the symptom checklist. These 10 items are 5-point Likert-type scales characterizing the time spent on compulsions (item 1), interference from obsessions (item 2), distress associated with obsessions (item 3), resistance to obsessions (item 4), subject's control over obsessions (item 5) and equivalent items for compulsions (items 6-10). The Y-BOCS has shown good psychometric properties and sensitivity to the therapeutic effects of medication and psychotherapy[26,202–206]. However, several problems have been identified for this scale, including a poor conceptual fit of the "resistance to obsessions" item, possibly contributing towards inconsistent factor structure, with some studies finding a two-factor (obsessions and compulsions) and others a three-factor structure (obsessions, compulsions and resistance to obsessions), as well as low sensitivity to change in severe cases and poor divergent validity relative to depressive symptoms[32,33,35,38,207,208].

To address some of these problems a revised version, the Y-BOCS-II, was published in 2000[38], with several differences relative to the original scale. Specifically, the obsessions and compulsions checklists are not formally subdivided into different symptom groups, some items in the symptom checklist were reworded and expanded, and a new checklist for avoidance was created. Additionally, in the severity scale, the item assessing "resistance against obsessions" was replaced by an item of "obsessions-free interval", the scoring for each item was revised from 0-4 to to 0-5, and the order of assessment of items was changed. Furthermore, avoidance was considered in the definition of severity, namely for the items of interference from obsessions and interference from compulsions. Finally, the definitions of obsessions and compulsions were rephrased, with several ancillary items removed from the text. Y-BOCS-II has excellent psychometric properties, with strong internal consistency, high test-retest and interrater reliabilities and strong correlations with other clinician-rated measures of obsessive-compulsive symptom severity, namely the National Institute of Mental Health Global Obsessive Compulsive Scale (NIMH-GOCS), and only moderate correlations with measures of worry and depressive symptoms[38]. These authors also conducted an exploratory factor analysis, the results of which were consistent with the obsession and compulsion severity subscales. Thus, the Y-BOCS scales are typically considered the gold-standard instrument in assessing severity of obsessive-compulsive symptoms[201,209], with the Y-BOCS-II translated and validated for other languages other than English[39,40].

Further exploration of the psychometric properties of the Y-BOCS-II is pertinent to for several reasons. In fact, to our knowledge, criterion validity of this scale has not been previously tested by comparing OCD patients with control samples, such as healthy subjects or, most importantly, patients with other similar disorders. Such comparisons would be important to define a cut-off value, allowing clinicians to establish that obsessive or compulsive symptoms may reflect an OCD diagnosis, rather than symptoms of a mood or anxiety disorder (e.g. rumination in depressive disorders and fear or worries in anxiety disorders)[210]. Furthermore, the underlying factor structure of the Y-BOCS-II is still a matter of debate[209], with the original American and the Thai versions showing a two-factor structure, as described above, while the Italian version had a different factor structure, with distinct dimensions[38–40]. Finally, the temporal stability of the Y-BOCS-II has only been tested in short intervals (at most 2 weeks) and it is clinically relevant to understand temporal stability for longer periods[38,39,41].

Here we explored the psychometric properties of a culturally adapted Portuguese translation of the Y-BOCS-II (PYBOCS-II), including internal consistency, factor structure, test-retest reliability, convergent validity and divergent validity. Importantly, we focused on the scale’s criterion validity, through comparisons of total scores between patients with OCD and control subjects, including both healthy volunteers and patients with other mood and anxiety disorders, as defined by a gold-standard instrument for diagnosis of OCD.

2.3. Objectives

2.3.1. Explore the factor structure of the Y-BOCS-II severity scale

2.3.2. Analyze convergent and divergent validity of the Y-BOCS-II

2.3.3. Test criterion-validity of the Y-BOCS-II for the diagnosis of OCD in adults

2.4. Methods

2.4.1. Participants

The protocol was approved by the Ethics Committee of Champalimaud Centre for the Unknown and by the Ethics Committee of Centro Hospitalar Psiquiátrico de Lisboa. Eligibility was assessed in 223 participants, recruited in either of the two clinical settings. Patients with a clinical diagnosis of OCD (n=60) were referred to the study by attending psychiatrists and psychologists, while control patients with other psychiatric diagnoses (n=35) were selected randomly from the institutional databases at each institution. A convenience sample of 128 healthy community dwelling subjects was also recruited at each of the two institutions. Exclusion criteria for all samples were: acute medical illness, active neurological disease or clinically significant focal structural lesion of the central nervous system; acute episode of neuropsychiatric disease requiring hospitalization; history or clinical evidence of chronic psychosis, dementia, developmental disorders with low intelligence quotient or any other form of cognitive impairment; current substance or alcohol abuse or dependence; and illiteracy or otherwise not understanding the study's instructions. For all participants except those in the OCD sample, current diagnosis of OCD, as assessed by structured diagnostic interviews (OCD subscale of the Structured Clinical Interview for the DSM-IV and MINI Neuropsychiatry Interview), was also an exclusion criterion. For the healthy volunteer sample, current or past history of any psychiatric disorder, as assessed by the MINI Neuropsychiatry Interview, was an additional exclusion criterion. Among the 223 participants that were assessed, 52 OCD patients, 18 patients with non-OCD mood or anxiety disorders and 117 healthy participants were eligible for the study.

2.4.2. Measures

2.4.2.1. Y-BOCS-II

The Y-BOCS-II consists of two main components: a 67-item symptom checklist and a 10-item severity-scale[38]. In the symptom checklist, 29 items assess the presence of specific obsessions, another 29 items assess the presence of specific compulsions, and the remaining nine items assess the presence of avoidance. Each item is dichotomously rated for current (i.e., within the past month) and past presence. In the severity scale, items assess, for the previous week, time spent with either obsessions or compulsions (items 1 and 6 respectively), obsession-free interval (item 2), resistance to compulsions (item 7), degree of control over either obsessions or compulsions (items 3 and 8 respectively), distress associated either with obsessions or with the impossibility of performing compulsions (items 4 and 9 respectively), and interference from either obsessions or compulsions (items 5 and 10 respectively). Avoidance items are considered to assess severity for either item 5, for avoidance related obsessions, or item 10 for avoidance related to compulsions. Each of the 10 items is rated in a 6-point scale (0-5) and 2 subscales are typically considered: an obsessions subscale (items 1-5) and a compulsions subscale (items 6-10). A more detailed description of the scale is given in the Introduction of this Chapter (section 2.2).

The YBOCS-II was not previously validated for use in adult populations speaking European Portuguese. To guarantee that linguistic and semantic equivalence of the Y-BOCS-II was preserved for use in such populations, we used a 3-step translation/back-translation method to obtain a Portuguese YBOCS-II (PYBOCS-II). For the first step, multiple independent translations from US English into European Portuguese, performed separately by four bilingual experts in Psychology or Psychiatry of Portuguese dominant language, were obtained, and then joined into a single consensus translation by the 4 translators. In the second step, back-translation of the consensus Portuguese translation back into English was performed by two bilingual translators, of English dominant language, that were not involved in the original translation. This was followed by comparison of the back-translated versions by the original translation team, for creation of a consensus back-translation.

In the last step, the consensus back-translation was compared with the original version by the initial translation team, and also sent for review and comments by the original authors of the Y-BOCS-II. This allowed for adjustments of the consensus Portuguese translation, to obtain a refined consensus Portuguese translation of the Y-BOCS-II. This version was then discussed among a panel of Portuguese-speaking experts in the fields of Psychiatry or Psychology, including but not limited to the original translation team, for assessment of face validity and proposal of additional adjustments for cultural adaptation. Finally, the scale was applied to a group of 10 patients suffering from OCD, followed by interviews for qualitative assessment of duration, cognitive effort and adequate comprehension of items. Considering the input from these patients, the translation was further adapted, and the final version of the PY-BOCS was defined.

2.4.2.2. Structured Clinical Interview for the DSM-IV, OCD Subscale (SCID-OCD)

The OCD Subscale of the SCID-IV is a semi-structured interview that allows for the diagnosis of current OCD according to DSM-IV criteria[211]. It has been validated for Brazilian Portuguese by Del-Ben and colleagues[212] and we adapted this version for European Portuguese. The SCID-OCD was used to discriminate between participants with and without OCD, for the purpose of criterion validity assessment.

2.4.2.3. MINI Neuropsychiatric Interview

The MINI is a brief structured clinical interview divided into 15 modules[213]. It allows for detection of major depressive disorder (MDD), dysthymia, suicide risk, manic and hypomanic episode, panic disorder, agoraphobia, social phobia, generalized anxiety disorder (GAD), OCD, post-traumatic stress-disorder, alcohol abuse or dependence, substance abuse or dependence, psychotic disorders, anorexia nervosa and bulimia nervosa, based on the rapid screening of DSM-IV diagnostic criteria. The interview has been translated to European Portuguese by Guterres, Levy and Amorim[214]. We used this version of the MINI to assess comorbidity and identify exclusion criteria.

2.4.2.4. Beck Depression Inventory II (BDI-II)

The BDI-II is a 21-item self-report screening instrument that assesses the presence of depressive symptoms in the previous 15 days[215]. Responses are scored from 0 (‘absent’) to 3 (‘severe’). It was validated to the Portuguese adult population by Campos and Gonçalves[216]. We used the BDI-II results to assess divergent validity with the PY-BOCS-II

2.4.2.5. State-Trait Anxiety Inventory (STAI)

The STAI is a widely-used 40 item self-report screening instrument that assesses the presence of anxiety symptoms[217]. It is composed of two subscales: the STAI-state and the STAI-trait. Trait anxiety corresponds to feelings of tension, apprehension and increased autonomic activity and is a relatively stable personality trait[217,218]. People with high trait anxiety have a tendency to perceive more situations as dangerous or threatening than people who have lower trait anxiety scores. State anxiety, on the other hand, fluctuates over time according to the presence of stressors. Individuals with high trait anxiety scores also tend to have higher state anxiety scores[217,218]. The scale was validated for use in Portuguese-speaking adults by Santos and Silva[219].

2.4.2.6. Coimbra Obsessive Inventory (COI - Inventário Obsessivo de Coimbra)

The COI is a self-report scale, developed for the Portuguese population, that assesses obsessive and compulsive symptoms through 12 dimensions, namely doubt and indecision, intrusive thoughts and covert rituals, magical thinking, slowness and repetition, need for control, need for order and symmetry, collection and hoarding, religious obsessions and compulsions, somatic obsessions, and obsessive and aggressive impulses[220]. It is subdivided in “frequency” and “emotional distress” subscales. The COI score was used to assess convergent validity for the PY-BOCS-II.

2.4.3. Procedures

In the non-blinded sample, after participants had responded to a global clinical questionnaire, instruments were applied in the following order: MINI, SCID-IV, PY-BOCS-II, BDI, STAI, COI. In the blinded sample, in a first session participants responded to the clinical questionnaire and the following instruments were applied, in the same order: MINI, SCID-IV, BDI, STAI, COI. In a second session, conducted by another researcher who did not have access to the first results, PY-BOCS-II was applied. Temporal stability was tested in a subsample of 27 OCD patients and 72 healthy participants by applying PY-BOCS-II a second time, four weeks after initial testing.

2.4.4. Data analysis

Descriptive statistics were calculated for sociodemographic and psychometric data, including means and standard deviations, minimum and maximum absolute values and percentage. We used independent samples t-tests to compare means between groups, except for gender (in which chi-square was used), with two-tailed significance values and the alpha-level was set to 0.05. We assessed several psychometric properties of the PY-BOCS-II. To estimate reliability, we analyzed internal consistency using Cronbach’s α and temporal stability using Pearson’s correlation coefficient. To assess dimensionality, exploratory factor analysis (EFA) with principal axis factoring and oblique rotation was performed in the Severity Scale. Factor analysis of the Symptom Checklist was not performed due to insufficient sample size of the OCD sample (a sample size of 5 to 10 participants per item is generally recommended - for 67 items a much larger sample size would be needed)[221]. To assess construct validity, we used Pearson’s correlation coefficient of PY-BOCS-II scores with COI scores for convergent validity, and with BDI scores and STAI scores for divergent validity. Finally, criterion validity was analyzed by studying the relationship between PY-BOCS-II scores and SCID-OCD classification using receiving operator characteristic (ROC) curves. Such curves are obtained by plotting the true positive rate (i.e. sensitivity) in function of the false positive rate (1-specificity), with each point in the curve representing a sensitivity/specificity pair corresponding to each possible decision threshold. Here, the area under the curve (AUC) of the ROC curve reflects the probability that randomly chosen individual with OCD had a higher PY-BOCS-II score than a randomly chosen individual without OCD diagnosis (with diagnosis defined by the SCID-OCD) The decision threshold, or cut-off value, for OCD diagnosis was then chosen according to the ROC curve as the total score that maximized sensitivity and specificity over all possible values.

In the non-blinded sample, after participants had responded to a global clinical questionnaire, instruments were applied in the following order: MINI, SCID-IV, PY-BOCS-II, BDI, STAI, COI. In the blinded sample, in a first session participants responded to the clinical questionnaire and the following instruments were applied, in the same order: MINI, SCID-IV, BDI, STAI, COI. In a second session, conducted by another researcher who did not have access to the first results, PY-BOCS-II was applied. Temporal stability was tested in a subsample of 27 OCD patients and 72 healthy participants by applying PY-BOCS-II a second time, four weeks after initial testing.

2.5. Results

2.5.1. Descriptive statistics

Sociodemographic data and mean scores of all psychometric instruments are presented in Table 1. While the control group was slightly younger than the OCD group, there were no significant differences in gender or education between samples. In the OCD sample, the most common comorbid diagnoses were MDD (38%), GAD (17%), prior MDD (15%), panic disorder (15%) and social phobia (12%). In the mood and anxiety disorders sample, the diagnoses were MDD (61%), GAD (39%), prior MDD (28%), past manic or hypomanic episode (22%), panic disorder (17%) and dysthymia (11%).

Table 1. Sociodemographic and psychometric data from each sample.

	OCD (n=52)		Non-OCD (n=135)		
	Range	Mean (SD)	Range	Mean (SD)	p-value
Gender (% male)	42.30%		30%		0.1
Age (years)	19-62	40.0 (10.0)	20-64	32.9 (9.5)	<0.001
Education (years)	7-23	14.7 (3.4)	4-23	15.4 (3.3)	0.2
Y-BOCS-II total score	0-45	22.7 (10.4)	0-25	1.8 (3.9)	<0.001
BDI total score	1-45	22.2 (13.6)	0-42	6.2 (9.1)	<0.001
STAI-state score	22-75	47.9 (14.9)	20-75	33.8 (10.8)	<0.001
STAI-trait score	26-77	56.9 (14.4)	20-74	32.8 (10.7)	<0.001
COI total score	18-332	137.9 (82.7)	0-290	31.9 (40.0)	<0.001

For all variables, mean and standard deviation are shown, except for gender (presented as percentage of males). Differences were tested using chi-square for gender and independent samples t-test for the other variables (p-values displayed). OCD, Obsessive-compulsive disorder; Y-BOCS-II, Yale-Brown Obsessive-Compulsive Scale-II; BDI-II, Beck Depression Inventory-II; STAI, State-Trait Anxiety Inventory; COI, Coimbra Obsessive Inventory.

Descriptive statistics of individual PY-BOCS-II Severity Scale items in the OCD sample are presented in Table 2. The PY-BOCS-II total score had a weak positive correlation with age (r=.28) when considering all participants, but in OCD patients this correlation was non-significant. Also, across all participants, there were no statistically significant differences between genders in any of the psychometric measures (*t*<1.23; *p*>0.21), and the correlations with education were either non-significant (for the PY-BOCS-II total score) or weak (r<0.3 for all other psychometric measures).

Table 2. Individual Y-BOCS-II item summaries for the OCD sample.

Items		Percentage of endorsement										
	Mean (SD)	0	1	2	3	4	5	Total	Sk	Ku	Item-total	α if deleted
1	2.0 (1.3)	9.6	31	31	17	3.8	7.7	100	0.7	0.2	0.6	0.9
2	2.4 (1.5)	12	21	21	19	23	5.8	100	0	-1	0.6	0.9
3	2.6 (1.6)	15	12	15	23	23	12	100	-0	-1	0.6	0.9
4	2.2 (1.3)	9.6	21	27	29	7.7	5.8	100	0.2	-0	0.5	0.9
5	1.8 (1.4)	19	27	27	12	12	3.8	100	0.5	-1	0.6	0.9
6	1.9 (1.3)	14	31	25	23	3.8	3.8	100	0.5	-0	0.6	0.9
7	2.5 (1.7)	21	12	12	19	25	12	100	-0	-1	0.5	0.9
8	2.8 (1.6)	14	5.8	19	19	31	12	100	-1	-1	0.7	0.9
9	2.7 (1.4)	7.7	14	25	25	17	12	100	-0	-1	0.7	0.9
10	1.8 (1.5)	27	19	17	23	12	1.9	100	0.2	-1	0.7	0.9

For each item of the Y-BOCS-II, the mean, standard deviation and the percentage of endorsement for each possible item score (range 0-5) is displayed. OCD, Obsessive-compulsive disorder; Y-BOCS-II, Yale-Brown Obsessive-Compulsive Scale-II; SD, Standard deviation; comp, compulsions; Sk, Skewness; Ku, Kurtosis; Item-total corr, Item-total correlation; α if deleted, Cronbach's α if item is deleted.

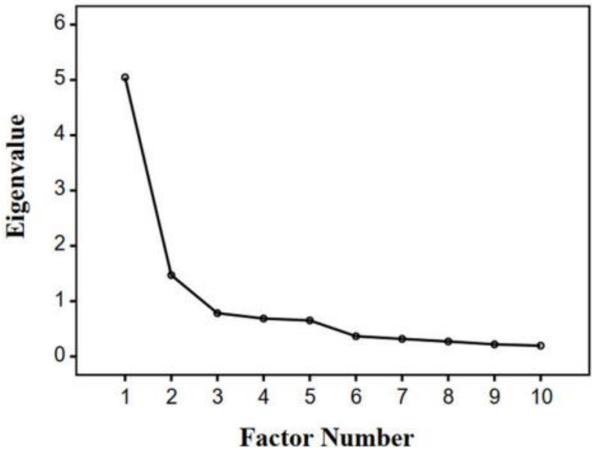
2.5.2. Reliability

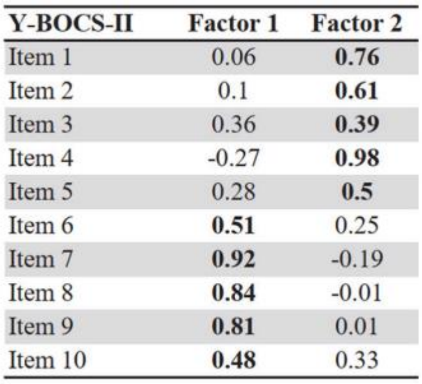
A Cronbach's alpha of 0.96 was obtained for the PY-BOCS-II severity scale, demonstrating robust internal consistency. A slightly lower value (0.94) was found for both the Obsessive subscale and Compulsive subscales, when tested separately. Furthermore, Cronbach's alpha remained stable with removal of any item from the scale (0.96 for all items), and corrected item-total correlations were in the range between 0.8 (item 4) and 0.87 (item 9).

Regarding temporal stability, assessed in 99 participants in the global sample, a Pearson's r of 0.94 (p<0.001) was obtained for the correlation of PY-BOCS-II total score at the first application and the same score 30 days later. When considering only the OCD sample (n=27), test-retest reliability was slightly higher (r=0.95, p<0.001). Finally, the temporal stability of the Obsessive subscale was higher than the temporal stability of the Compulsive subscale, both when considering all participants (r=0.94 vs. r=0.89 respectively) or only OCD patients (r=0.92 vs. r=0.84 respectively).

2.5.3. Dimensionality

We conducted exploratory factor analysis using principal axis factoring with promax rotation in the OCD sample. The Kaiser-Meier-Olkin measure of sample adequacy was 0.836, above the recommended value of 0.6, and the Bartlett's test of sphericity was significant ($X^2_{(45)}$ = 265.75, p<0.001). Two factors with eigenvalues > 1 were obtained (eigenvalues of 5.05 for the first factor and 1.47 for the second factor) and this two-factor solution was consistent with the deflection of the scree plot (Fig. 11). The pattern matrix revealed that items 6-10 had higher loadings on factor 1 (all > 0.4) and items 1-5 on factor 2. Item 3 had relatively small loadings on both factors, although with slightly higher loading on factor 2. The correlation between factor 1 and factor 2 was 0.55.





Y-BOCS-II	Factor 1	Factor 2
Item 1	0.06	**0.76**
Item 2	0.1	**0.61**
Item 3	0.36	**0.39**
Item 4	-0.27	**0.98**
Item 5	0.28	**0.5**
Item 6	**0.51**	0.25
Item 7	**0.92**	-0.19
Item 8	**0.84**	-0.01
Item 9	**0.81**	0.01
Item 10	**0.48**	0.33

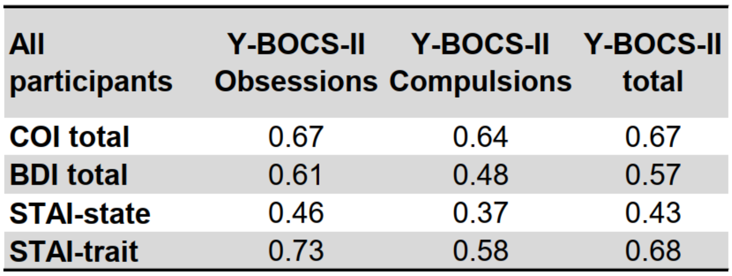
Figure 11. Scree plot (exploratory factor analysis) and pattern matrix for Y-BOCS-II factors in the OCD sample. In the pattern matrix, standardized weights of a regression

analysis in which item responses are predicted from their levels of the underlying factors are represented. Factor loadings above 0.4 or highest factor loading shown in bold.

2.5.4. Construct Validity

Measures for construct validity, using correlations between the PY-BOCS-II and several self-report psychometric measures, are shown in Table 3. For convergent validity, we found a significant and strong correlation between the PY-BOCS-II total score and the score for a self-report obsessive-compulsive inventory (COI) r=0.67, p<0.001), with similar correlations with each of the COI subscales (r=0.67 for the frequency subscale and r = 0.66 for the emotional distress subscale, both with p<0.001). For divergent validity, the correlations between the Y-BOCS-II total score and the STAI-state scores was only moderate (r=0.43, p<0.001), higher, but still moderate, for the BDI score (r=0.57, p<0.001), and strong for the STAI-trait (r=0.68, p<0.001). Furthermore, the PY-BOCS-II Compulsions subscale had lower correlation with BDI and both STAI scores than the Obsessions subscale (Table 3) suggesting a better divergent validity for the Compulsions subscale.

Table 3. Correlations between psychometric measures and Y-BOCS-II partial and total score in all participants.



All participants	Y-BOCS-II Obsessions	Y-BOCS-II Compulsions	Y-BOCS-II total
COI total	0.67	0.64	0.67
BDI total	0.61	0.48	0.57
STAI-state	0.46	0.37	0.43
STAI-trait	0.73	0.58	0.68

Pearson's product moment correlation coefficient used as correlation measure. All correlations are highly significant (p's < 0.001). Y-BOCS-II, Yale-Brown Obsessive-Compulsive Scale-II; COI, Coimbra Obsessive Inventory; BDI-II, Beck Depression Inventory-II; STAI, State-Trait Anxiety Inventory

2.5.5. Criterion validity

To assess criterion validity, we created Receiver Operating Characteristic (ROC) curves (Fig. 12), using the SCID-OCD as the discriminator between participants with OCD (n=35) and controls (n=135; Fig. 12 left panel). An area under the curve (AUC) of 0.96 (95% Confidence interval [95% CI]: 0.92, 0.99) was obtained, and further analyses of the ROC curve values showed that a PY-BOCS-II total score of 13 points, when used as a cut-off for diagnosis, correctly identifies OCD with a sensitivity of 85% and a specificity of 97% (table 4).

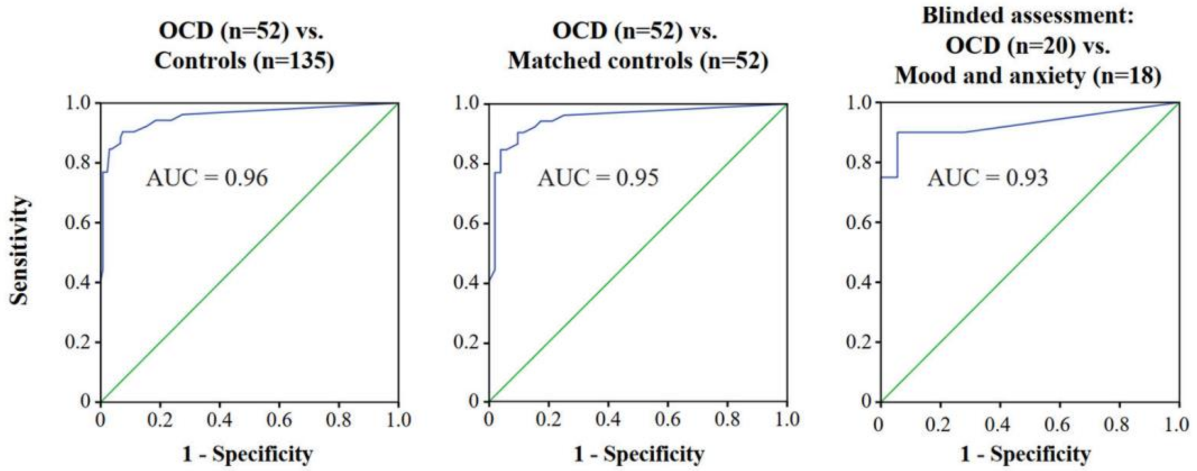


Figure 12. ROC curves for use of the Y-BOCS-II to identify OCD.

Plot of the true positive rate (1—specificity) against the false positive rate (sensitivity) for the different possible cut-offs of the Y-BOCS-II using the SCID-OCD as the diagnostic instrument. In the left panel, all participants were considered. In the middle panel, OCD and age-, gender-, and education-matched controls (balanced mixture of healthy subjects and patients with mood and anxiety disorders) are considered. In the right panel, patients who completed a blinded assessment are considered. ROC, Receiver operating characteristic; Y-BOCS-II, Yale-Brown Obsessive-Compulsive Scale-II; OCD, Obsessive-compulsive disorder; AUC, Area under the curve.

Table 4. Coordinates for the ROC curve of the Y-BOCS-II using all participants

Y-BOCS-II cut-off score	Sensitivity	Specificity
0.5	96.20%	72.60%
1.5	94.20%	76.30%
2.5	94.20%	78.50%
3.5	94.20%	81.50%
4.5	92.30%	84.40%
5.5	90.40%	88.90%
6.5	90.40%	91.10%
7.5	90.40%	91.90%
8.5	90.40%	92.60%
9.5	88.50%	93.30%
10.5	86.50%	93.30%
11.5	84.60%	96.30%
13	84.60%	97.00%
14.5	76.90%	97.80%
15.5	76.90%	98.50%
16.5	76.90%	99.30%
17.5	75.00%	99.30%
18.5	73.10%	99.30%
19.5	67.30%	99.30%
20.5	63.50%	99.30%
21.5	59.60%	99.30%
22.5	55.80%	99.30%
23.5	50.00%	99.30%
24.5	44.20%	99.30%
25.5	40.40%	100.00%

To further explore the discriminatory capacity of the Y-BOCS-II, a similar analysis was performed comparing the OCD sample with a group of age-, gender- and education matched controls (frequency-matched balanced mixture of healthy subjects and patients with mood and anxiety disorders; Fig. 12 middle panel). The AUC was similar (AUC=0.95; 95% CI: 0.91, 0.99) and a total score cut-off of 13 points remained optimal, with sensitivity of 85% and specificity of 96%. Importantly, the same analyses were repeated in data from a subgroup of patients with either OCD (n=20) or other mood and anxiety disorders (n=18), for whom PY-BOCS-II was applied by a researcher blinded to diagnosis and to the results of other psychometric tests. In this group (Fig. 12 right panel), AUC was only slightly lower (AUC=0.93; 95% CI: 0.84, 1) and the 13-point cut-off resulted in sensitivity of 90%, and specificity of 94%, in diagnosis of OCD.

2.6. Discussion

Here, we have translated and successfully validated the Y-BOCS-II for the Portuguese adult population. A translated and culturally adapted version of the scale had excellent reliability and was valid for assessment of the severity of obsessive-compulsive symptoms. Our results further supported a two-factor structure for the scale, consistent with the obsessions and compulsions subscales proposed by the original authors. Importantly, and addressing the main objective of this study, we have demonstrated, to the best of our knowledge for the first time, that the YBOCS-II adequately discriminates between OCD and non-OCD patients, and that a cut-off of 13 points for the Y-BOCS-II total score has excellent sensitivity and specificity.

Our results on reliability of the PY-BOCS-II are in line with the studies that have previously assessed the psychometric properties of this scale. Storch and colleagues found strong internal consistency (Cronbach's alpha = 0.89), similar to what was described later for the Thai (0.94) and Italian (0.83) versions of the scale[38–40]. Regarding test-retest reliability, high values were reported in the original description of the psychometric properties of the scale (Intraclass correlation [ICC] > 0.85), as well as for the Italian version (ICC = 0.74), while the Thai version did not assess this psychometric dimension[38–40]. Recently, psychometric properties of the original American version of the Y-BOCS-II were retested, with findings of good internal consistency (Cronbach's alpha = 0.86), acceptable test-retest reliability (r = 0.64-0.81) and excellent inter-rater reliability (ICC 0.97-0.99)[41]. Our findings for internal consistency (Cronbach's alpha = 0.96) and test-retest reliability (r = 0.94-0.95) are in the upper range of prior studies, suggesting that the process of translation and cultural adaptation was successful. Furthermore, other authors have suggested that temporal stability be tested with longer test-retest intervals than 2 weeks[38,39,41]. Ours is, to our knowledge, the first study to demonstrate stability of test scores after 4 weeks.

Regarding dimensionality, and due to lack of consensus regarding the factor structure of the Y-BOCS-II, we decided to perform an exploratory factor analysis (EFA) rather than a confirmatory factory analysis (CFA), as was common in previous studies.

While, in general terms, our results replicate previous findings of a two-factor solution corresponding to obsessions and compulsions, there a few subtle but noteworthy differences[38,40]. Specifically, for the original and Thai versions of the task, interference from obsessions (item 5) had high loadings (>.4) on both factors, with the authors of the Thai version also reporting higher loadings of distress associated with obsessions (item 4) on the compulsions factor than the obsessions factor[40]. Loadings in our data were more clearly distributed between the two factors, with the first five items mainly loading on a factor that is consistent with an Obsessive dimension, and the last five items loading mainly on the second factor, consistent with a Compulsive dimension. Unexpectedly, item 3 ("control over obsessions") loaded similarly on both factors, possibly because a subset of patients may feel that their level of control over obsessions is dependent on the frequency and severity of compulsions. Importantly, our results are in marked contrast with those for the Italian version of the scale, which revealed a "symptom severity" factor (items 1-4 and 6-9) and "interference from symptoms in daily life" factor (items 5 and 10)[39]. It is unclear whether these differences in factor structure reflect true cultural differences across different countries with respect to the presentation of OCD, or are merely due to methodological differences, namely regarding sample size.

With regards to convergent validity, the PY-BOCS-II showed a correlation of 0.67 with self-reported obsessive-compulsive symptom scores in the COI. This correlation was observed even though a high score in the COI reflects a high number of different symptoms causing distress, but not necessarily the severity of individual symptoms[26,222], while the Y-BOCS-II measures severity of OCD symptoms regardless of the number of different symptoms. Other authors have found low to moderate correlations between Y-BOCS-II scores and scores on self-reported OCD symptom assessment tools such as the Obsessive-Compulsive Inventory-Revised (OCI-R), while correlations with clinician-rated obsessive-compulsive symptom scales such as the National Institute of Mental Health Global Obsessive Compulsive Scale are stronger (e.g., r=.85)[38]. Assessing convergent validity against a clinician-rated scale would thus, in all likelihood, have yielded a more robust correlation for the P-Y-BOCS-II. For divergent validity, the PY-BOCS-II total score showed a moderate correlation with both depression and state-anxiety scores, and a strong correlation with trait-anxiety scores. This observation replicates the findings of previous studies on the psychometric properties of the Y-BOCS, that also found weak correlations with self-reported measure of anxiety and moderate to strong correlations with self-reported measures of depression such as the Inventory of Depressive Symptomatology – Self-Report (r=.35)[38], the Patient Healthy Questionnaire (r=.45)[40], the BDI (r=.40)[39]. , or the Depression Anxiety Stress Scale – Depression subscale (r=.41)[41]. Together, the currently available data suggests that divergent validity regarding depression symptoms is, at best, only moderate. This was also a problem with the first edition of the Y-BOCS, and may be related to the high co-morbidity between OCD and major depressive disorder (MDD), which may be as high as 50%[223–225]. As to the robust correlation between the PY-BOCS-II and STAI-trait anxiety, it may simply reflect the fact that patients with more severe OCD tend to have higher levels of longstanding comorbid anxiety, rather than a true limitation in the scale's ability to discriminate between these two dimensions

Our findings of higher correlations with self-reported depression and anxiety symptoms in the Obsessive subscale than in the Compulsion subscale suggest that the latter may have better divergent validity. This finding is in line with the results from Storch and colleagues. In their study, the Y-BOCS-II Compulsion subscale had higher correlations with the NIMH-GOCS and with the OCI-R and lower correlations with the PSWQ and with the IDS-SR when compared with the Obsessions subscale[38]. For the Thai version of the Y-BOCS-II, the correlations of subscales with depressive symptoms were non-significant and in the Italian version such were not presented[39,40]. This finding is particularly interesting because it could suggest higher tendency for obsessions than for compulsions in patients with comorbid OCD and MDD and higher tendency for compulsions in patients with OCD only.

The main objective of this chapter, however, was to clarify criterion validity for this scale. AUC of the ROC curves demonstrated that the Y-BOCS is accurate in discriminating between patients with OCD and other without the disorder. To our knowledge, this is the first study exploring criterion-related validity of the Y-BOCS-II in OCD patients, healthy controls and in patients with other psychiatric disorders. The cut-off value that we propose (13) is in line with previous findings using the first edition of the Y-BOCS[206]. Using that first version, Farris and colleagues have shown that a posttreatment YBOCS score of ≤ 14 was the best predictor of symptom remission and that a posttreatment YBOCS score of ≤ 12 was the best predictor of wellness (defined as symptom remission, good quality of life and high level of adaptive functioning)[206]. However, it is important to note that this study focused on treatment response and that while the first edition of the YBOCS has an upper limit of 40 points, the upper limit of Y-BOCS-II is 50 points. Nonetheless, the cut-off we propose can be useful from a diagnostic perspective, because clinicians often assess patients with obsession-like ideas or compulsive-like behaviors, which may or may not suffer from OCD.

Nevertheless, our study is not free of limitations. Regarding validation of the PY-BOCS-II, information about inter-rater reliability would be reassuring. However, all previous studies of psychometric measures of the Y-BOCS-II which have performed this analysis have found excellent inter-rater reliability[38,39,41]. Furthermore, it would have been desirable to have a control group without significant differences in age, particularly considering the weak positive correlations with age across all psychometric instruments used. However, the Y-BOCS-II had the weakest correlations with age and, in the OCD group, the correlation between Y-BOCS-II and age was non-significant. Nevertheless, to eliminate any potential effects of such differences in the ROC curves, we selected a sample of age-, gender- and education-matched controls and repeated our main analysis only with this group, obtaining confirmation of our previous results. Also, in a subsample of individuals (32 OCD participants), raters were not blind to diagnosis, which could lead to criterion contamination. To account for this potential limitation, we also created ROC curves using only the subset of OCD and non-OCD patients that were assessed in a blinded fashion. While the number of participants included in this analysis was lower, the results obtained were very similar to the remaining ROC curves thus validating our findings. In the future it could be important to repeat this specific analysis using larger OCD and non-OCD clinical samples. The use of the SCID-OCD as a diagnostic instrument can also be considered a limitation because it has never been validated for the Portuguese population. However, it has been validated for Brazilian Portuguese and the adaptation to European Portuguese was very straightforward. It is also important to acknowledge that Hoarding Disorder, which was part of OCD in DSM-IV, is considered a separate disorder in DSM-5. However, none of the participants included in this study presented exclusively hoarding symptoms (Y-BOCS Symptom Checklist items 26 and 46). Finally, future studies could address the properties of the Y-BOCS-II regarding classification of treatment sensitivity, as has been done for the first version of the scale.

In conclusion, we have successfully translated and validated the Y-BOCS-II for use in the Portuguese adult population, showing that the Portuguese version of the Y-BOCS-II maintains the psychometric properties of the original version in evaluating the severity of obsessive-compulsive symptoms. Using this version of the task we have also, for the first time, assessed criterion validity of the Y-BOCS-II, by exploring the its capacity to distinguish between patients with OCD and subjects in several clinical and non-clinical groups, using both a blinded and a non-blinded design. Our results suggest that a Y-BOCS-II total score cut-off of 13 has good sensitivity and excellent specificity in identifying OCD. However, we also found problems in divergent validity with symptoms of depression and anxiety, as previously described by others.

Chapter 3. Application of a new sequential decision task in healthy and clinical populations

3.1. Abstract

Multi-step decision tasks have emerged as an influential tool to investigate reinforcement learning (RL) in humans, and to assess associations between RL and certain behavioral disorders, such as obsessive-compulsive disorder (OCD), where deficits in model-based RL have been described. Prior to performing these tasks, subjects typically receive detailed information about task structure. Thus, it remains unclear how different RL systems contribute when subjects learn exclusively from experience, and how explicit information about task structure modifies RL strategy. To address these questions, we created a two-step task requiring minimal prior instruction, and assessed performance both prior to and after providing explicit information on task structure, in healthy volunteers, patients with OCD and patients with other mood and anxiety disorders. Initially, model-free control dominated, with model-based control emerging only in a minority of subjects after significant task experience, and not at all in patients with OCD or patients with mood and anxiety disorders. However, once explicit information about task structure was provided, a dramatic increase in the use of model-based RL was observed, similarly across healthy volunteers and both patient groups, including OCD. Importantly, only in patients with OCD, model-free strength RL increased with uninstructed experience and was not significantly changed by the debriefing, contrary to the remaining groups. Additionally, in all groups, after instructions, model-free action value updates were influenced more by state values and less by trial outcomes, and, in healthy volunteers, subject choices became more perseverative, consistent with changes in exploration strategy. Our results suggest that, in domains where humans lack prior knowledge, model-free RL predominates in sequential decision problems, particularly among patients with OCD, and that given specific task information, model-based RL emerges in healthy volunteers as well as patients with several psychiatric disorders, including OCD.

3.2. Introduction

Within psychology and neuroscience, it is thought that the brain uses multiple systems to choose which actions to perform[125,129,132,170,226,227]. One widely held distinction is made between goal-directed and habitual actions, the former mediated by predictions of the specific outcomes of each action, and the latter induced automatically by specific stimuli or states[123–126]. This cognitive and behavioral classification of actions is thought to correspond, at least in part, to a computational distinction between two different types of reinforcement learning (RL) algorithms designated respectively as model-based and model-free[132,167,170,181]. Model-based RL learns the state-action-state transition structure of the environment, allowing behavioral trajectories to be simulated. The value, i.e. long run utility, of different actions can be evaluated online using this model in conjunction with the learned immediate rewards available in each state. This allows for statistically efficient use of experience and behavioral flexibility, at the cost of the computational demands of planning. Model-free RL by contrast, maintains trial-and-error estimates of the value of each state or action, and updates these based on reward prediction errors. Model-free RL supports rapid action selection at low computational cost, but uses experience less efficiently, hence taking longer to adapt to changes in the environment. It is thought that the brain takes advantage of the complementary strengths of both approaches through metacognitive mechanisms which estimate whether the payoff for more accurate prediction is worth the computational costs of planning[170,183,228].

Multi-step decision tasks have emerged as a powerful approach to identify model-based and model-free RL in humans[180–183]. In such tasks, subjects move through a sequence of states to obtain rewards, typically with non-stationary reward or transition probabilities, forcing ongoing learning. The contribution of model-based and model-free RL are assessed by looking at the granular pattern of how subjects update their choices in light of recent experience. The most popular such task is the ‘two-step’ task, employing a choice between two actions which lead probabilistically to two states where rewards may be obtained[181]. Each action commonly leads to one of the states but, on a minority of trials, may alternatively lead to the state commonly reached by the other action. In this task, model-based and model-free RL are identified according to how the trial outcome (rewarded or not) and state transition (common or rare) interact to affect the subsequent choice. Under model-free control, the agent will tend to repeat first-step choices that are followed by reward, irrespective of the state transition. In contrast, under model-based control, while the agent will behave similarly after common state transitions, the opposite behavior should be observed following rare transitions, i.e., changing the first-step choice if a reward was obtained and repeating the first-step choice if a reward was not obtained. The two-step task has been used to study neural correlates of model-based and model-free control in healthy subjects[181,187,188,190–192,229–234], and to investigate decision making in clinical populations[185,186,235–237], typically integrating the balance between model-based and model-free control into a single weighting parameter.

As in most human decision tasks, subjects performing the two-step task receive extensive prior instruction about task structure. Although there is an extensive prior literature showing that instruction profoundly shapes human behavior in operant[193–195] and fear[196,197] conditioning, as well as value-based decision making[198–200], instruction effects have not been explored in multi-step tasks where model-based and model-free control can be dissociated. It therefore remains unclear how these different RL mechanisms contribute to action selection in situations where subjects must learn task structure directly from experience, and how providing explicit information about task structure modifies each system.

To address these questions, we created a modified version of the two-step task, requiring minimal prior instruction, that was initially applied with no prior information about the task state space, transition structure, or reward probabilities, and then repeated following debriefing about these elements of the task structure. Behavior in this task was tested in healthy volunteers as well as in a sample of patients with OCD, previously reported to have deficits in model-based RL[235,237] (Fig. 12), to control for the effects of psychotropic medication and the effects of unspecific mood and anxiety symptoms, in a sample of patients with other mood and anxiety disorders.

3.3. Objectives

3.3.1. Develop a new sequential decision task which allows to isolate instructed and uninstructed RL strategies

3.3.2. Explore the effects of experience and the effects of explicit knowledge on model-based and model-free control in healthy subjects

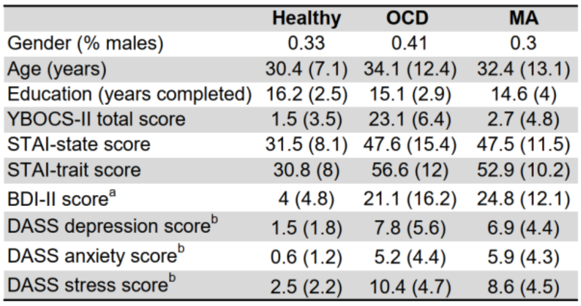
3.3.3. Assess model-based and model-free RL strategies in OCD patients in uninstructed and instructed sequential action choice

3.4. Results

3.4.1. Model-based control develops with task experience

One-hundred and eight healthy volunteers were recruited both in Lisbon and NY to characterize behavior in the reduced two-step task. Sociodemographic and psychometric data from these individuals are shown in table 5.

Table 5. Sociodemographic data and results from psychometric scales and neuropsychological tests.



	Healthy	OCD	MA
Gender (% males)	0.33	0.41	0.3
Age (years)	30.4 (7.1)	34.1 (12.4)	32.4 (13.1)
Education (years completed)	16.2 (2.5)	15.1 (2.9)	14.6 (4)
YBOCS-II total score	1.5 (3.5)	23.1 (6.4)	2.7 (4.8)
STAI-state score	31.5 (8.1)	47.6 (15.4)	47.5 (11.5)
STAI-trait score	30.8 (8)	56.6 (12)	52.9 (10.2)
BDI-II score[a]	4 (4.8)	21.1 (16.2)	24.8 (12.1)
DASS depression score[b]	1.5 (1.8)	7.8 (5.6)	6.9 (4.4)
DASS anxiety score[b]	0.6 (1.2)	5.2 (4.4)	5.9 (4.3)
DASS stress score[b]	2.5 (2.2)	10.4 (4.7)	8.6 (4.5)

[a] only in Lisbon sample; [b] only in New York sample; YBOCS-II = Yale-Brown Obsessive-Compulsive Scale-II; STAI = State-Trait Anxiety Inventory; BDI-II = Beck Depression Inventory-II

Two slightly different versions of the task were tested as described in figure 13. Healthy volunteers in Lisbon were randomized between the Fixed version (n=40) and the Changing version (n=42), while in NY only the former was applied (n=27). In either version of the task, participants performed 1200 trials on a single day, divided in 4 sessions of 300 trials each. Analysis of behavior at reward probability side-reversals, namely the pre-reversal fraction of correct first-step choices and exponential fit of post-reversal first-step choice change (see Fig. 13C), were used to assess overall task performance (see methods for details). Assessment of reinforcement learning strategy (model-based vs. model-free), however, was performed according to stay-probability analysis, in addition to fitting reinforcement learning models to observed behavior. For stay-probability analysis, the effect of events on one trial, in particular the outcome (rewarded vs. non-rewarded) and the transition probability from first-step action to second step state (common vs. rare; see Fig. 13B), on the subsequent first-step choice are quantified (see methods for details). In the original two-step task, since transition probabilities are fixed and also explained and demonstrated to participants before the task starts, it is typically assumed that estimates of these transition probabilities are not updated online in response to the experienced state transitions[181]. In the current task, this is somewhat modified, since participants have no prior information about transition probabilities and must learn them online from experienced state transitions. Model-based RL is classically associated with an interaction between transition and outcome, because a model-based agent understands that outcomes following rare transitions should primarily influence the value of the alternate first-step choice, commonly leading to the state where that outcome was received[181,238]. Furthermore, we have previously shown that, in simulations of the behavior of a model-based agent, when transition probability estimates are updated based on experienced state transitions, a multivariate analysis of stay probability demonstrates loading on a transition parameter (i.e. common transitions promote repetition of the same choice) in addition to the transition-outcome interaction parameter[238].

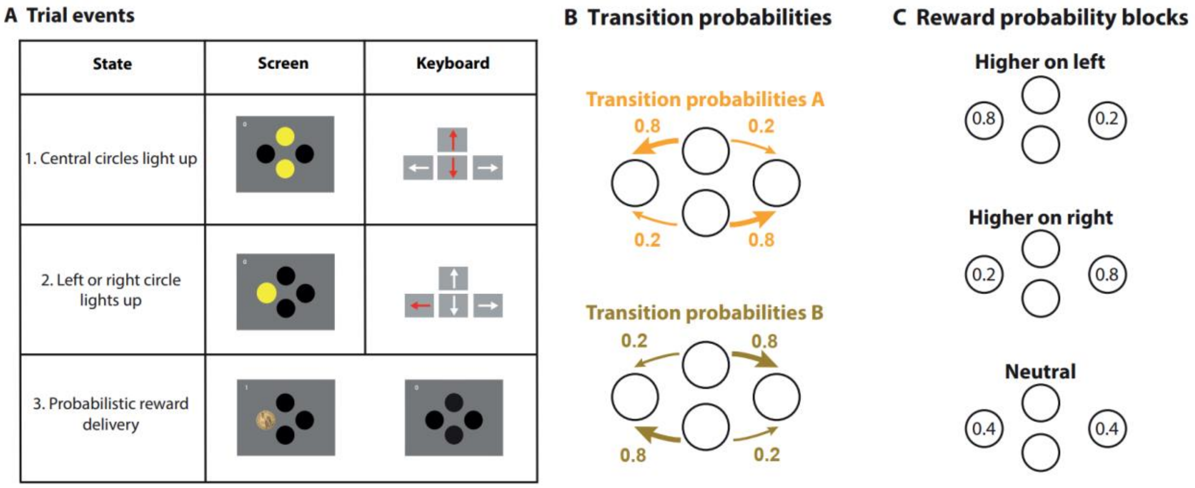

A Trial events
State
Screen
Keyboard
1. Central circles light up
2. Left or right circle lights up
3. Probabilistic reward delivery
B Transition probabilities
Transition probabilities A
0.8
0.2
0.2
0.8
Transition probabilities B
0.2
0.8
0.8
0.2
C Reward probability blocks
Higher on left
0.8
0.2
Higher on right
0.2
0.8
Neutral
0.4
0.4

Figure 13. Reduced two-step task structure.

The task was performed on a computer interface with 4 circles visible on a grey screen: 2 central circles (upper and lower) and two side circles (right and left). Each circle was colored yellow when available for selection, and black when unavailable, and could be selected by pressing the corresponding arrow key (up, down, left or right) on the computer keyboard. **A)** Trial events: Each trial started with the central circles turning yellow, prompting the choice between either upper or lower circle (a1). This choice caused one of the side circles (left or right) to become yellow (a2), with differing probabilities (see b). The subject should then select the yellow side circle resulting in probabilistic monetary reward. A reward was indicated by the side circle changing to the image of a coin (a3) while no reward was indicated by the circle changing back to black colour (see C). **B)** Transition probabilities: At each trial, choosing one circle (e.g., upper) commonly [p=0.8] lit up one side circle and rarely [p=0.2] the other side circle, with inverse probabilities for choosing the alternate circle (e.g., lower). The transition probabilities were fixed (either A or B, counterbalanced across subjects) in the Fixed version of the task, or underwent reversals from A to B in the Changing version (see methods for details). **C)** Reward probability blocks: The reward probabilities (p) upon selection of the side circles changed in blocks that were either neutral [p=0.4 for both left and right sides], or higher on one of the sides (p=0.8 vs p=0.2, i.e., non-neutral blocks), Non-neutral blocks ended when subjects consistently chose the first-step option (upper or lower) that most frequently lead to the high reward probability side. Neutral blocks ended probabilistically, independent of subjects' behavior (see methods for details). To maximize reward rate subjects must choose the first step action which commonly leads to the second-step state with higher reward probability, tracking the best option across reward-probability reversals.

We first assessed learning effects in the task with fixed transition probabilities (Fixed task) by comparing behavior between the first and third sessions (Fig. 14). Overall task performance, as assessed by reward probability reversal analysis, while indicative of appropriate performance, did not improve significantly between sessions 1 and 3, regarding both the pre-reversal fraction of correct choices (session 1: 0.74, session 3: 0.72, *P*=0.63) and the exponential fit time constant of the post-reversal first-step choice change (session 1: 24.8, session 3: 15.8, *P*=0.16; Fig. 14A). However, stay probability analysis revealed change between sessions 1 and 3 (Fig. 14B), with an increased influence of both transition (common vs rare; *P*=0.004) and transition-outcome interaction (*P*=0.002), but only a trend towards evidence of a change in the influence of outcome (rewarded vs. non-rewarded; *P*=0.06) (Fig. 14 C). These changes are consistent with increased influence of model-based RL on decision-making, from sessions 1 to 3. Importantly, we also found a significant correlation between use of model-based RL, as assessed by loading on the transition–outcome interaction parameter across sessions 1 to 3, and the number of rewards obtained in the same sessions (rho=0.7, *P*<0.001; Fig. 14D).

Fitting of reinforcement learning (RL) models to observed behavior across sessions 1 to 3 was then performed, Model-comparison indicated that a mixture model including model-free and model-based components fit the data better than a purely model-free or purely model-based model, as reflected by lower Bayesian Information Criteria (BIC) scores (Fig. 14E, left panel). Furthermore, models which included a “bias” parameter, capturing bias towards the upper or lower first-step choice, and a “perseveration” parameter, capturing a tendency to repeat the previous choice, fit the data better than a model not including these parameters (Fig. 14E, right panel). We also compared parameter values of the fitted models for sessions 1 and 3 (Fig. 14F). The value learning rate increased significantly between session 1 and 3 (*P*=0.04), but no other parameters changed significantly. The discrepancy with increased influence of model-based RL across time as assessed by stay-probability analysis likely reflects lower statistical power to detect strategy changes in the strongly non-linear and higher parameter count RL models. Heterogeneity of behavior across subjects, as is evident in the very wide range of value learning rates exhibited across subjects, likely also reduces the ability of the RL model to capture changes due to learning.

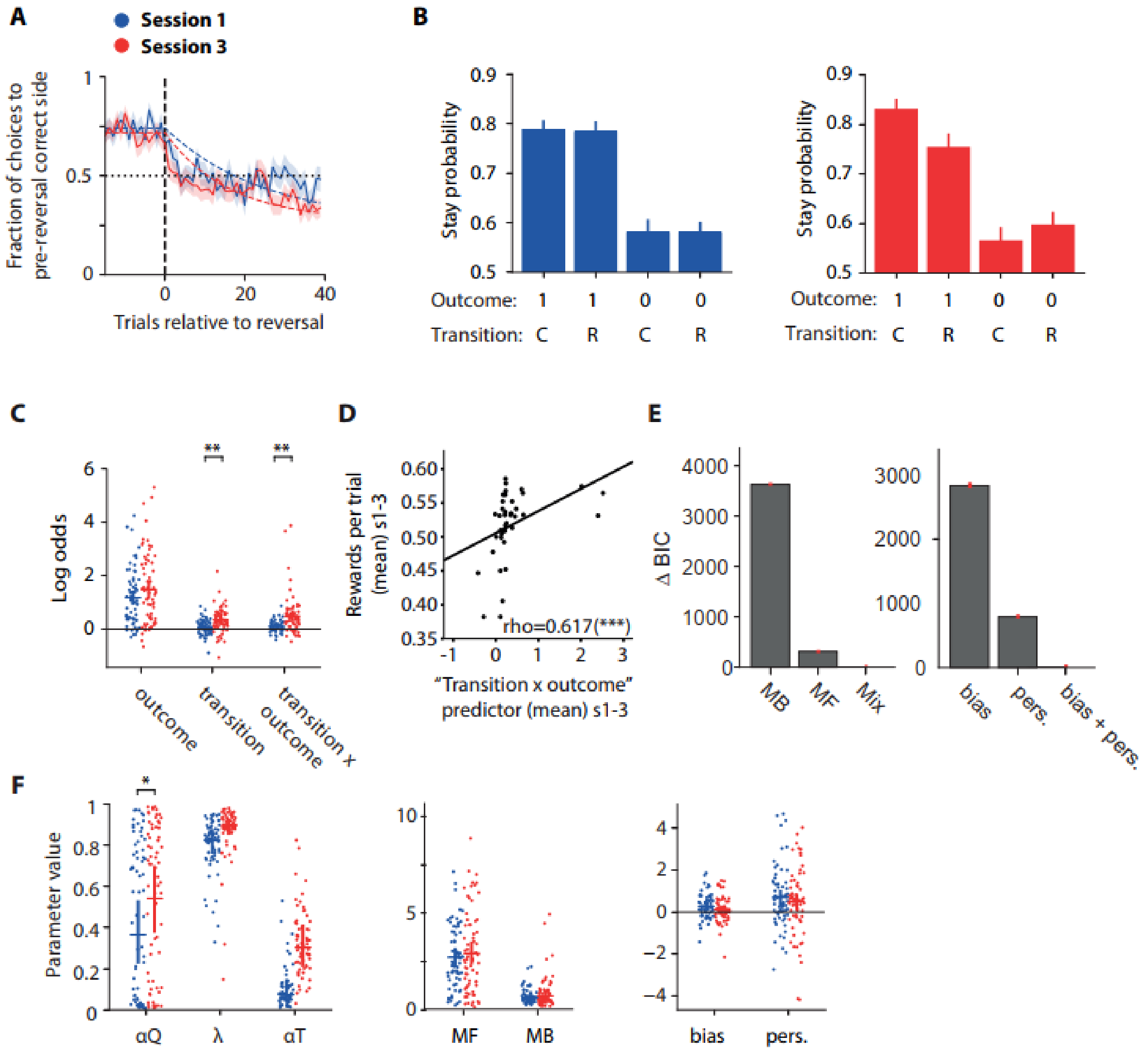


Figure 14. Learning effects in the fixed (transition probabilities) version of the reduced two-step task.

A) First-step choice trajectories around reversals. In this and other panels, blue indicates session 1 while red indicates session 3. Solid lines show cross-subject mean trajectory, while dashed lines show exponential fits to the average trajectories. Confidence regions (mean ± across subject standard error) are represented by shaded areas. **B)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicate across subject standard error (SEM). The left panel shows data from session 1, the right panel from session 3. **C)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Positive loading on the 'outcome' predictor indicates a tendency to repeat rewarded choices, while the 'transition' predictor reflects a tendency to repeat choices followed by common transitions, and the 'transition x outcome' interaction predictor indicates a tendency to repeat choices that were rewarded following a common transition, or that were not rewarded following a rare transition. Dots indicate maximum a posteriori parameter values for individual subjects, while bars indicate the population mean and 95% confidence interval of the mean. Statistical significance of differences in factor loadings for each predictor between session 1 (blue) and 3 (red) was assessed using permutation tests. **D)** Scatter plot showing the relationship between model-based RL across sessions 1, 2 and 3, as captured by mean loading in the “transition x outcome” predictor, and the mean number of rewards per trial in the same sessions. Correlations were assessed using Spearman’s rank correlation coefficient. **E)** Bayesian Information Criteria (BIC) model comparison for sessions 1-3. Left panel, comparison of model-based (MB), model-free (MF) and MB-MF mixture (Mix) models. Right panel, comparison of mixture model with bias parameter, perseveration parameter, and bias + perseveration parameters. **F)** Comparison of mixture model fits between session 1 and session 3. Dots and bars are represented as in panel C. RL model parameters: MF: Model-free strength, MB: Model-based strength, αQ: Value learning rate, λ: Eligibility trace, αT: Transition prob. learning rate, bias: Choice bias, pers.: Choice perseveration. In all figures significant differences are indicated as: * *P*<0.05, ** *P*<0.01, *** *P*<0.001.

In the version of the task where the transition probabilities underwent reversals (Changing task, Fig. 15), behavior also revealed appropriate overall task performance that did not improve between sessions 1 and 3, regarding both the pre-reversal fraction of correct choices (session 1: 0.77, session 3: 0.76, *P*=0.98) and the exponential fit time constant of the post-reversal first-step choice change (session 1: 22.8, session 3: 14.0, *P*=0.3; Fig. 15A). However, stay-probability analysis did not show increased influence of transition and transition-outcome interaction as observed in the Fixed task (*P*=0.2 for both), but rather an increased influence of trial outcome (*P*=0.01), previously associated with a model-free direct reinforcement strategy[181,238](Fig. 15B,C). In accordance with results in the Fixed task, there was a significant correlation between loading on the transition–outcome interaction parameter and the number of rewards obtained (rho=0.4, *P*<0.01; Fig. 15D). Model comparison indicated that the mixture and model-free RL models fitted the data much better than the model-based-only model, and that the difference in BIC scores between the mixture model and model-free-only model was negligible (ΔiBIC = 3; Fig. 15E left panel). Similarly to the Fixed task, the model including both “bias” and “perseveration” parameters fit the data better than a model lacking these parameters (Fig. 15E, right panel). For consistency with analyses of the Fixed task, we used the mixture model to look for differences in behavior between sessions 1 and 3, but found no significant change in model parameters (Fig. 15F).

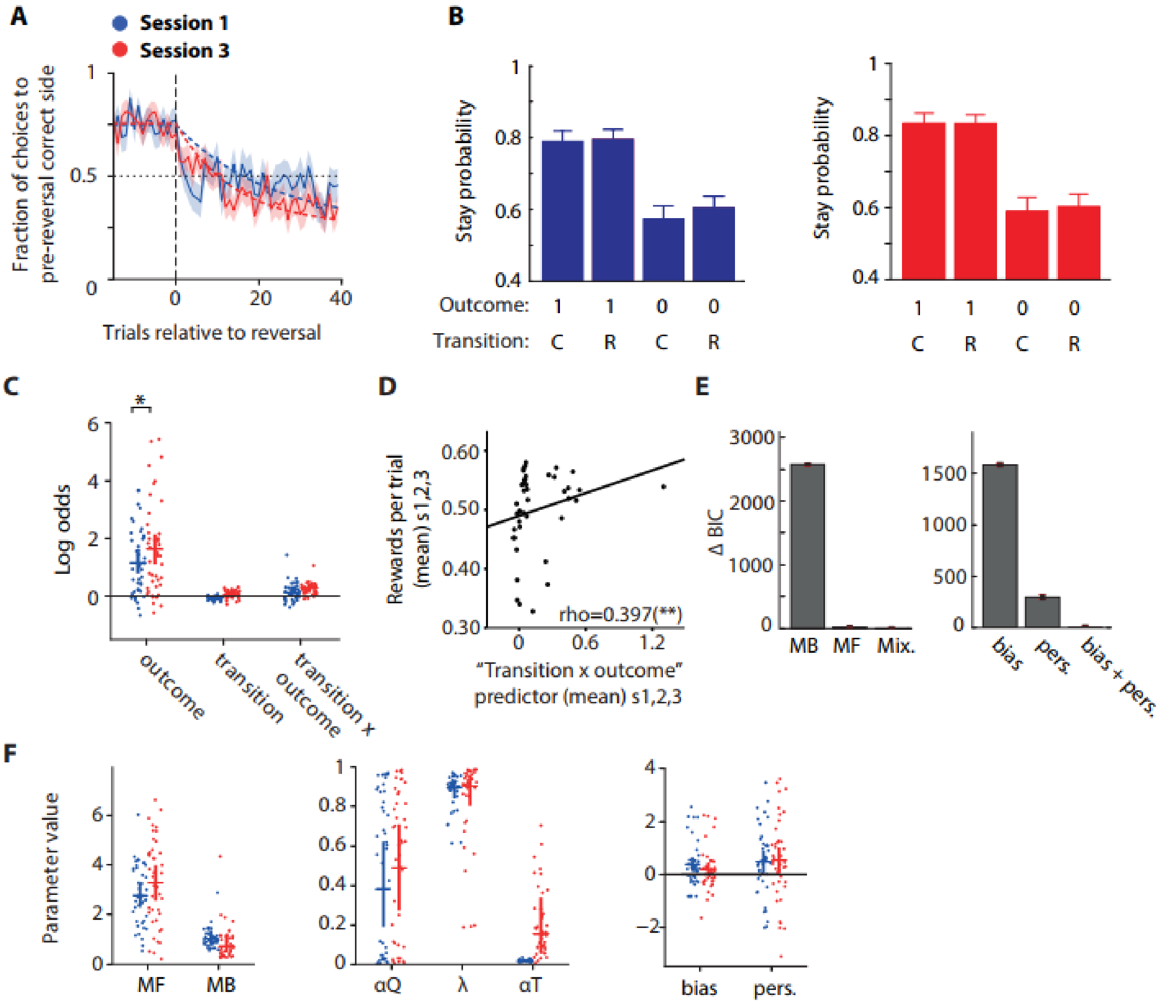


Figure 15. Learning effects in the changing (transition probabilities) version of the reduced two-step task.

A) Mean first-step choice trajectories around reversals. In this and all panels, blue indicates session 1 while red indicates session 3. Dashed lines show exponential curves fitted to the average trajectories to obtain estimates of the time-course of learning following reversals. Confidence regions (mean ± cross subject standard error) are represented by shaded areas. **B)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicate the cross subject standard error (SEM). The left panel shows data from the first session, the right panel shows data from session 3. **C)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Positive loading on the 'outcome' predictor indicates a tendency to repeat rewarded choices. Positive loading on the 'transition' predictor reflects a tendency to repeat choices followed by common transitions. Positive loading on the 'transition x outcome' interaction predictor indicates a tendency to repeat choices that were rewarded following a common transition, or that were not rewarded following a rare transition. Dots indicate maximum a posteriori loadings for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. Statistical significance of differences in factor loadings for each predictor between session 1 (blue) and 3 (red) were evaluated using permutation tests. **D)** Scatter plot showing the relationship between model-based RL (captured by mean loading in the "transition x outcome" predictor across sessions 1, 2 and 3) and the mean number of rewards per trial (in the same sessions). Correlations were assessed using Spearman's rank correlation coefficient. **E)** Bayesian Information Criteria (BIC) model comparison for sessions 1-3. Left panel, comparison of model-based (MB), model-free (MF) and mixture (MF+MB) models. Right panel, comparison of mixture model with bias parameter, perseveration parameter, and bias + perseveration parameters. **F)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: αQ: Value learning rate, λ: Eligibility trace, αT: Transition prob. learning rate, MF: Model-free strength, MB: Model-based strength, bias: Choice bias, pers.: Choice perseveration.

3.4.2. Explicit knowledge affects both model-based and model-free control.

We next assessed how providing explicit information about the task's structure changed subjects' behavior, by comparing behavior in sessions 3 and 4 both in a group that did and a group that did not receive debriefing about task structure after session 3. To avoid ceiling effects in subjects who already understood the task well, these analyses only included subjects for whom a likelihood ratio test indicated model-based RL was not being used significantly in session 3 (57 of 67 subjects for the Fixed task; Fig. 16A, F).

In the Fixed task, debriefing dramatically increased the number of subjects identified by a likelihood-ratio test as using model-based RL in session 4 (21/41 subjects in debriefing group, 1/16 subjects in no-debriefing group; z = 3.13, *P*=0.002, z-test for difference in proportions; Fig. 16A, F). Subjects in the debriefing group adapted faster to reversals following debriefing (*P*<0.001, Fig. 16B), an effect that was not found in the no-debriefing group (*P*=0.4, Fig. 16G), even though the session by group interaction did not reach significance (*P*=0.4). Debriefing also strongly affected how events on one trial influenced the subsequent choice (Fig. 16C, D, H, I), with increased influence of state transition (*P*<0.001; session by group interaction *P*=0.03) and transition-outcome interaction (*P*<0.001; session by group interaction *P*=0.01) on stay probability, consistent with increased use of model-based RL following the debriefing. RL mixture model fits of the pre and post debriefing data (Fig. 16E, J) indicated that, consistent with regression analysis, the influence of model-based action values on choice was increased by debriefing (*P*<0.001; session by group interaction *P*=0.01). Importantly, even though the session by group interaction was not significant for this parameter (*P*=0.7) debriefing also reduced model-free action values (P = 0.008). The RL model further indicated that other aspects of behavior were affected by the debriefing. Specifically, the eligibility trace parameter was decreased (*P*=0.004, session by group interaction *P*=0.03), such that updates of model-free first step action values depended less on the trial outcome (rewarded or not) and more on the value of the second-step state that was reached, while the perseveration parameter, assessing repetition of the previous choice, was increased (*P*<0.001, session by group interaction *P*=0.001). Post debriefing, value learning rates were also higher (P<0.001), but the session by group interaction was not significant (*P*=0.6). As all participants in the no debriefing group were recruited in Lisbon, as opposed to the debriefing group, which included both subjects recruited in Lisbon and subjects recruited in New York, we also ran the same analyses including only participants recruited in Lisbon and the results were maintained (data not shown). As expected, no significant differences were found for any of the comparisons between sessions 3 and 4 in the no debriefing group (Fig. 16G-J). We also ran the same analyses without excluding subjects who were model-based at sessions 3 and the results were similar, with the exception of the intra-individual debriefing effects in the model-free action values and in the value learning rate, which were not significant when all subjects were included (data not shown). These findings indicate that providing explicit knowledge of task structure not only promoted use of model-based RL but also affected value updates in the model-free system and value-independent choice preservation.

In the Changing version of the task, debriefing did not increase the use of model-based RL. The fraction of subjects identified as using a model-based strategy at session 4 was the same in the debriefing and no-debriefing groups (debriefing group 2/12, no-debriefing group 4/24; z = 0, *P*=1, z-test for difference of proportions; Fig. 17A, F). Subjects in the debriefing group adapted faster to reversals in session 4 than session 3 (*P*=0.03, Fig. 17B), and the logistic regression analysis showed an increased influence of the trial outcome on subsequent choice in session 4 compared to 3 in the debriefing group (*P*=0.02, Fig. 17D), but the session by group interaction did not reach significance in both cases (*P*= 0.2 and 0.06 respectively). The influence of the transition and transition-outcome interaction parameters on subsequent choice were unaffected by debriefing (P = 0.99 and 0.3 respectively, Fig. 17D) and no parameters of the RL model differed significantly pre and post-debriefing (P > 0.19, Fig. 17E). As expected, no significant differences were observed in any analyses between sessions 3 and 4 in the no-debriefing group (Fig. 17G-J). These results indicate that in the this more complex Changing task subjects either did not to understand the debriefing or decided the effort of trying to use information about the task structure was not worthwhile.

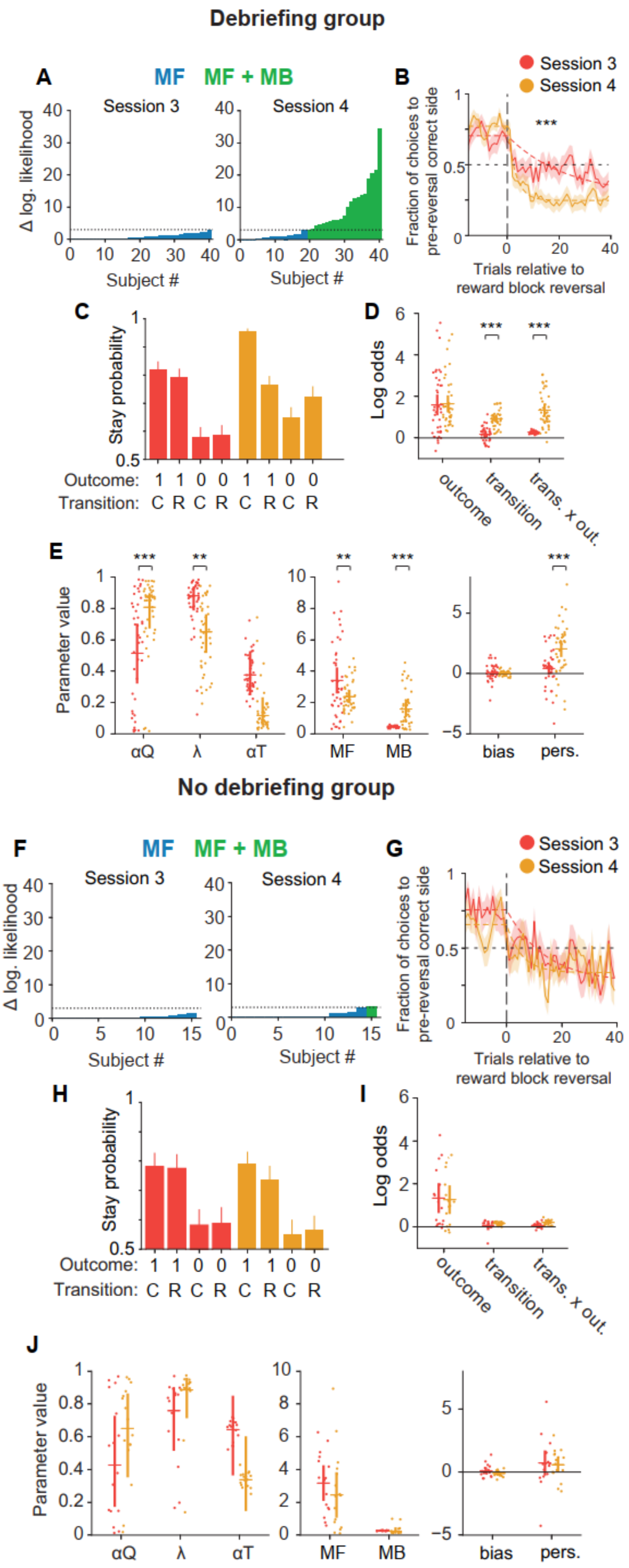


Figure 16. Effects of explicit knowledge in the fixed (transition probabilities) version of the reduced two-step task.

(A, F) Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). Data was analyzed separately for groups with (A) and without (F) debriefing. Y-axis shows difference in log likelihood between mixture (model-free + model-based) RL model and model-free only RL model. Blue bars indicate subjects for which likelihood ratio test favors model-free only model, green bars indicate subjects for which test favors mixture model, using a $p<0.05$ threshold for rejecting the simpler model. In these and other panels we compared sessions 3 and 4 only in the subjects for whom a likelihood ratio test indicated that model-based RL was not used in session 3. **(B, G)** Mean first-step choice trajectories around reversals. In these and remaining panels, red indicates session 3 (before instruction in the debriefing group) while yellow indicates session 4 (after instruction in the no-debriefing group). Solid lines show cross-subject mean trajectory. Dashed lines show exponential curves fitted to the average trajectories to obtain estimates of the adaptation time-course of learning following reversals. Confidence regions (mean ± across subject standard error) are represented by shaded areas. **(C, H)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicated the cross subject standard error of the mean (SEM). **(D, I)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Dots indicate maximum a posteriori parameter values for individual subjects, while bars indicate the population mean and 95% confidence interval on the mean. **(E, J)** Comparison of mixture model fits. Dots and bars are represented as in panels D and I. RL model parameters: MF: Model-free strength, MB: Model-based strength, αQ: Value learning rate, λ: Eligibility trace, αT: Transition prob. learning rate, bias: Choice bias, pers.: Choice perseveration.

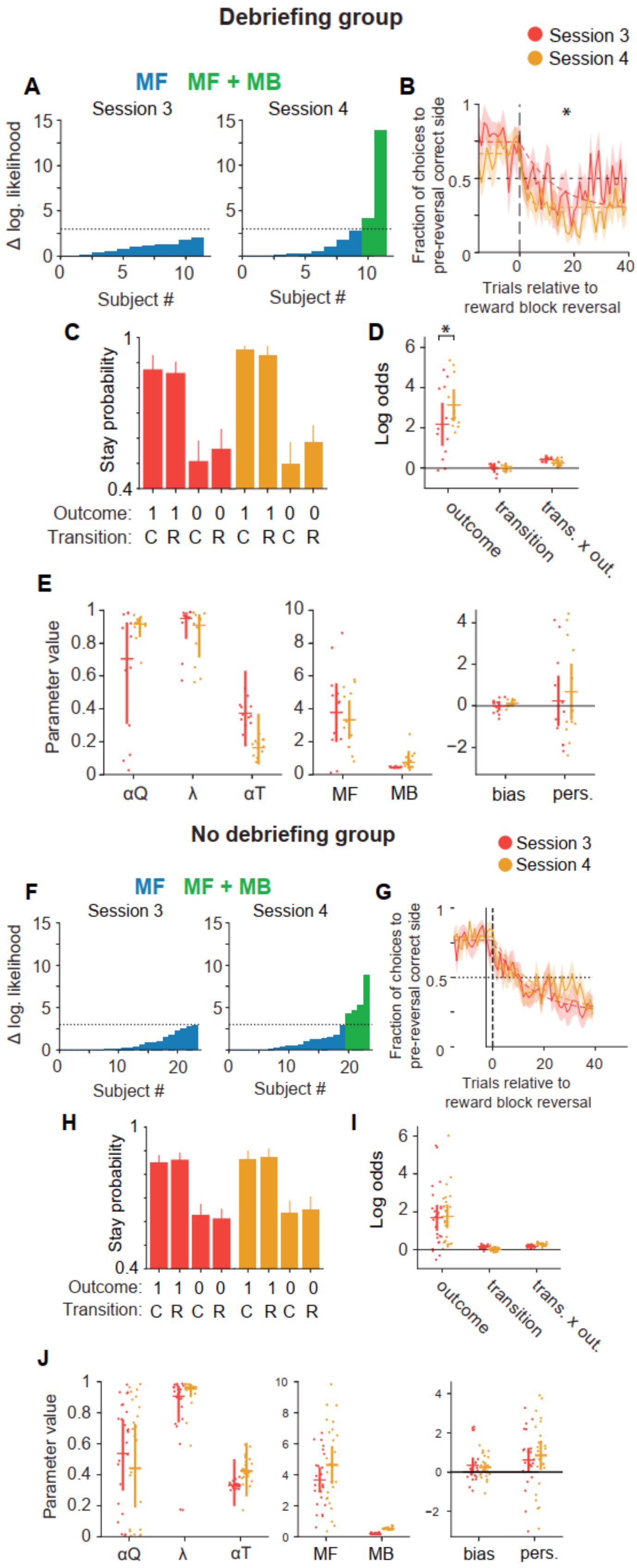


Figure 17. Effects of explicit knowledge in the changing (transition probabilities) version of the reduced two-step task.

(A, F) Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). Data was analyzed separately for groups with (A) and without (F) debriefing. Y-axis shows difference in log likelihood between mixture (model-free + model-based) RL model and model-free only RL model. Blue bars indicate subjects for which likelihood ratio test favors model-free only model, green bars indicate subjects for which test favors mixture model, using a p<0.05 threshold for rejecting the simpler model. We compared sessions 3 and 4 only in the subjects for whom a likelihood ratio test indicated that model-based RL was not used in session 3. **(B, G)** Mean first-step choice trajectories around reversals. In these and all remaining panels, red indicates session 3 (before instruction) while gold indicates session 4 (after instruction). Dashed lines show exponential curves fitted to the average trajectories to obtain estimates of the adaptation time-course of learning following reversals. Confidence regions (mean ± across subject standard error) are represented by shaded areas. **(C, H)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicated the cross subject standard error of the mean (SEM). In each group data was analyzed separately for session 3 (red graph) and session 4 (gold graph). **(D, I)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. **(E, J)** Comparison of mixture model fits between session 3 (red) and session 4 (gold) in the group without instruction (left panels) and the group with instruction (right panels). RL model parameters: αQ: Value learning rate, λ: Eligibility trace, αT: Transition prob. learning rate, MF: Model-free strength, MB: Model-based strength, bias: Choice bias, pers.: Choice perseveration.

3.4.3. Effects of experience and of knowledge in patients with OCD

Forty-six patients with OCD and 50 control patients with other mood and anxiety disorders completed the fixed version of the task, with debriefing between session 3 and session 4. Sociodemographic and psychometric data results from these samples are shown in Table 1. In the OCD group, the speed of adapting to reversals did not change between sessions 1 and 3 (*P*=0.6) but there was a trend towards an increased fraction of correct choices at the end of blocks (*P*=0.06, Fig. 18A). A session-by-group interaction test for changes in overall performance over learning between OCD subjects and healthy controls further indicated a trend towards significance for block end fraction correct choices (*P*=0.07), but no differences in the reversal time constant (*P*=0.5). In the analysis of stay probabilities, the OCD group did not show an increase in the influence of transition or transition-outcome interaction with learning that was seen in the controls (*P*=0.08 and *P*=0.71 respectively, Fig. 18B, C), but the influence of trial outcome increased over learning (*P*=0.001), which was not seen in the controls. However, the session by group interactions were significant only for the transition-outcome interaction parameter (*P*=0.01), but not for the transition (P = 0.6) or outcome parameters (*P*=0.2). Consistently, RL mixture model fits to sessions 1 and 3 (Fig. 18D) showed an increase in the influence of model-free action values on choice over learning (*P*=0.008) that was not seen in controls, with the session-by-group interaction revealing a trend towards significance (*P*=0.07). Furthermore, OCD patients did not show the increase in the value learning rate over learning seen in healthy volunteers (*P*=0.1; session-by-group interaction, P = 0.98). In the mood and anxiety subjects, we did not see any significant change in overall performance between session 1 and 3 as assessed by the reversal analysis (Fig. 18E). The logistic regression analysis of stay probabilities (Fig. 18F,G) showed an increased influence between session 1 and 3 of trial outcome (*P*<0.001) and transition (P = 0.01) on repeating choice, but not the transition-outcome interaction predictor (*P*=0.66). RL model fits showed only an increased learning rate for values (*P*=0.01). No session-by-group interactions reached significance for differences in learning effects from the control group. Overall, these data suggest a different pattern of learning from experience in patients with OCD, with a failure to learn the task-transition structure and exhibit model-based RL, but an increased influence of model-free action values, and hence trial outcome's direct influence on choice.

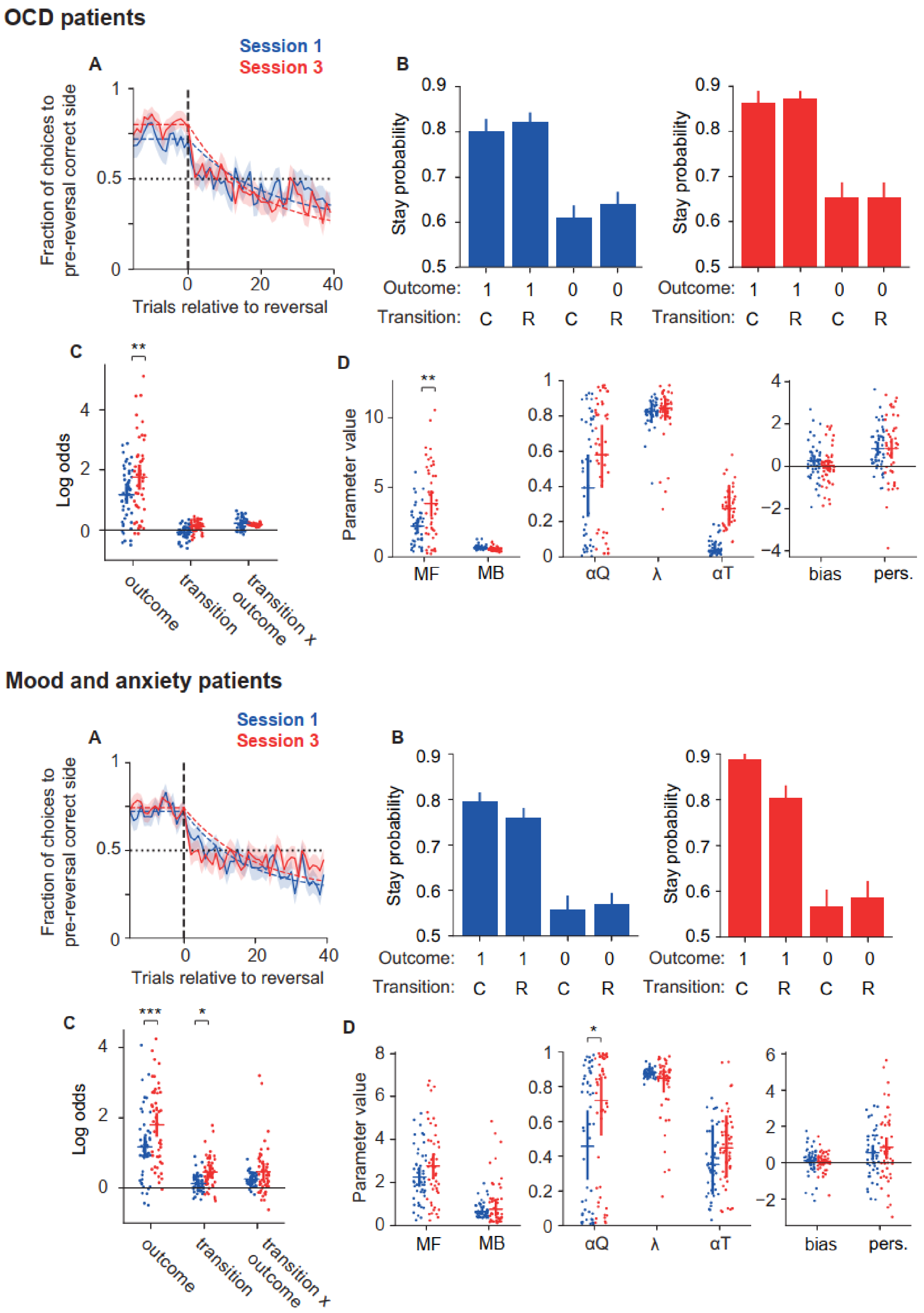


Figure 18. Learning effects in clinical samples

(A, E) Mean first-step choice trajectories around reversals. In all panels, blue indicates the first session while red indicates session 3. Solid lines show cross-subject mean trajectory.

Dashed lines show exponential curves fitted to the average trajectories to obtain estimates of the adaptation time-course of learning following reversals. Confidence regions (mean ± cross subject standard error) are represented by shaded areas. **(B, F)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicated the cross subject standard error of the mean (SEM). The left (blue) panel shows data from the first session, the right (red) panel shows data from session 3. **(C, G)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Dots indicate maximum a posteriori loading for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. **(D, H)** Comparison of mixture model fits. Dots and bars are represented as in panels C and G. RL model parameters: MF: Model-free strength, MB: Model-based strength, αQ: Value learning rate, λ: Eligibility trace, αT: Transition prob. learning rate, bias: Choice bias, pers.: Choice perseveration.

After debriefing, patients with OCD adapted faster to reversals (*P*=0.007, Fig. 19B). This improvement in performance was accompanied by an increase in both the transition (*P*=0.002) and transition-outcome interaction (*P*<0.001) parameters, as was found for healthy volunteers and consistently with increased use of model-based RL (Fig. 19C, D). This was confirmed by RL model fitting (Fig. 19E), which showed an increase in the influence of model-based action values on choice (*P*<0.001), while replicating the effects seen in healthy volunteers of reductions in the eligibility trace (*P*=0.01) and perseveration (*P*=0.02) parameters. However, contrary to what was found in healthy volunteers, debriefing did not reduce model-free action values, there was only a trend towards a reduction from sessions 3 to 4 (*P*=0.06). Significant session-by-group interaction was found only for increases in perseveration, that was larger in healthy volunteers (*P*=0.04). The effects of debriefing in patients in the mood and anxiety disorder group was similar. There were significant increases in the speed of adaption to reversals (*P*<0.001, Fig. 19G) as well as the influence of transition (*P*=0.002) and transition-outcome interaction (*P*<0.001) parameters on stay probability (Fig. 19H, I). Consistently, the RL model fit showed that debriefing increased the influence of model-based action values on choice (*P*<0.001). This group also replicated the debriefing effect on model-free action values (*P*=0.02), eligibility trace (*P*=0.04) and value learning rate (*P*=0.014) parameters observed in healthy volunteers (Fig. 19J) but, contrary to controls, there was no effect of debriefing on perseveration (*P*=0.4) and there was a significantly decreased learning rate for transitions (*P*=0.047). Significant session-by-group interaction was found only for the lack of effect on perseveration in the mood and anxiety patients (*P*=0.001 We also ran the session-by-group interaction analysis in the clinical vs. healthy group comparisons without excluding subjects who used model-based control at session 3, and the results were similar, except for the loss of significant interaction for the perseveration parameter in the OCD vs. healthy controls, which nevertheless remained at trend level (P=0.06).

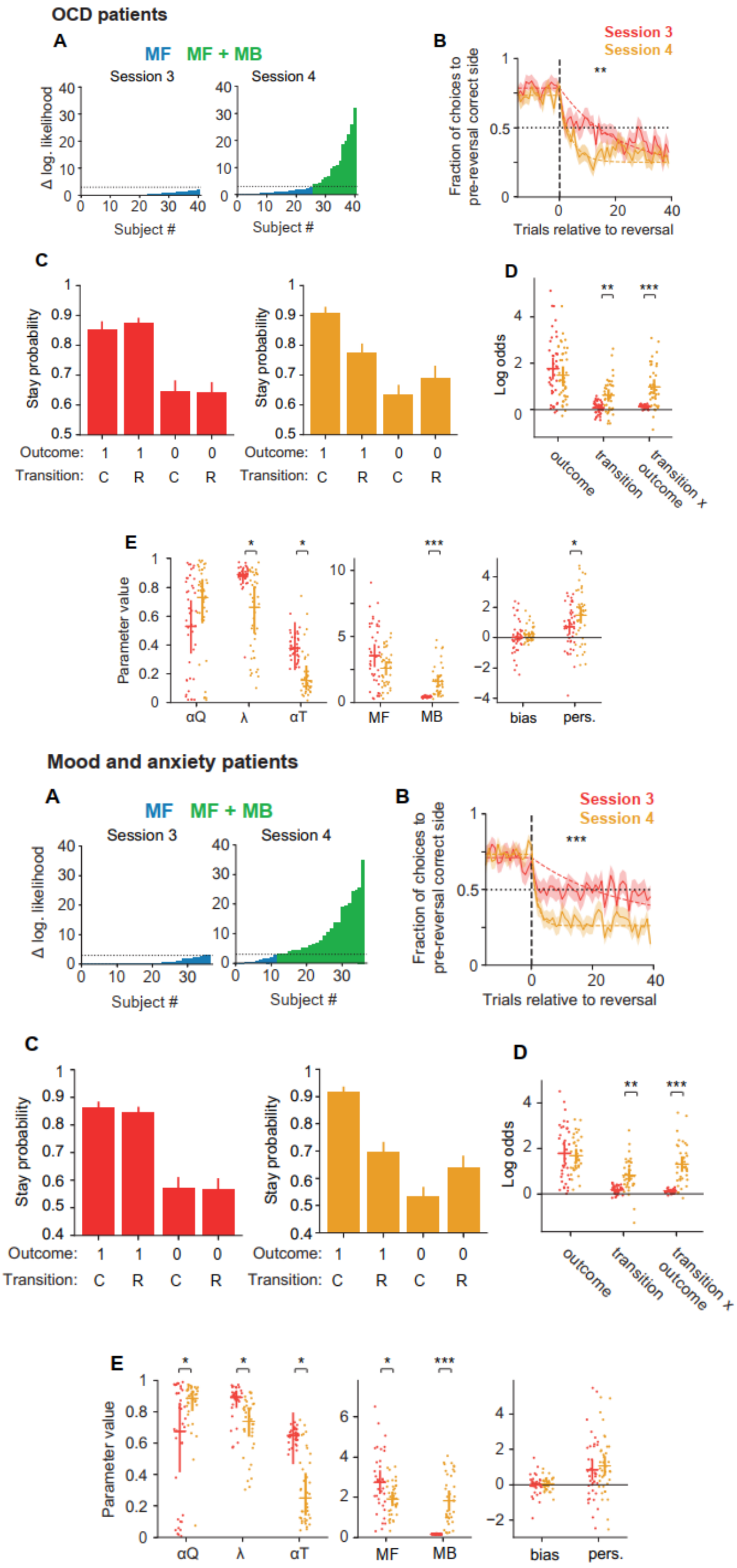


Figure 19. Debriefing effects in clinical samples.

(A, F) Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). Y-axis shows difference in log likelihood between mixture (model-free + model-based) RL model and model-free only RL model. Purple bars indicate subjects for which likelihood ratio test favors model-free only model, green bars indicate subjects for which test favors mixture model, using a p<0.05 threshold for rejecting the simpler model. We analyzed the effects of debriefing using only those subjects a likelihood ratio test indicated did not use model-based RL in session 3. **(B, G)** Mean first-step choice trajectories around reversals. In this and all panels, red indicates session 3 (before instruction) while yellow indicates session 4 (after instruction). Solid lines show cross-subject mean trajectory. Dashed lines show exponential curves fitted to the average trajectories to obtain estimates of the adaptation time-course of learning following reversals. Confidence regions (mean ± cross subject standard error) are represented by shaded areas. **(C, H)** Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicated the cross subject standard error of the mean (SEM). Data was analyzed separately for groups with (left panel) and without (right panel) instruction. **(D, I)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Dots indicate maximum a posteriori loading for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. **(E, J)** Comparison of mixture model fits. Dots and bars are represented as in panels C and G. RL model parameters: MF: Model-free strength, MB: Model-based strength, αQ: Value learning rate, λ: Eligibility trace, αT: Transition prob. learning rate, bias: Choice bias, pers.: Choice perseveration.

3.5. Discussion

We developed a novel two-step decision task in which participants received minimal pre-task instruction to examine how model-based and model-free RL contribute to behavior when task structure must be learned directly from experience, and how behavior is modified when information about task structure is subsequently provided. We found that, in an unfamiliar domain, participants were, overall, surprisingly poor at learning to use even simple forward models for action selection, but that delivery of instructions strongly increased the use of model-based control, among other diverse effects on behavior, such as changing how model-free value propagated through the task's state space, and modifying sampling strategies. Importantly, while OCD patients exhibited a deficit in model-based control when learning directly from experience, such deficits were no longer observed after delivery of explicit information on task structure, which is contrary to what has been described for other, more complex, two-step tasks[184], and raises questions about the nature of these deficits. Furthermore, equivalent deficits were found in patients with other mood and anxiety disorders, arguing against specificity for OCD. A pattern of enhanced model-free control, particularly during learning from experience, seemed however, much more specific for OCD.

Although model-free RL was the dominant strategy prior to instructions, the influence of model-based RL increased over learning. This contrasts with habit formation in instrumental conditioning, where actions are initially goal-directed, but become habitual with extended experience[152], which is thought to parallel a transition from model-based to model-free control[158] . This transition, and arbitration between model-based and model-free control more generally, is thought to occur through meta-cognitive mechanisms which assess whether the benefits of improved prediction accuracy are worth the costs of model-based evaluation[158,183,228]. The different trajectory of arbitration between model-based vs model-free arbitration in the current task likely reflects its dynamic nature, where ongoing changes in reward probability prevent the model-free system from converging to accurate value estimates, and hence dominating behavior late in learning. The more complex state space compared with more typical instrumental conditioning makes model learning more demanding, possibly due to uncertainty in early learning. In fact, it has been recently suggested that performance during initial stages of action selection tasks may be primarily based on trial-and-error exploration, with progression towards model-based RL occurring in intermediate stages, as subjects acquire a model of the environment[239]. In any case, it is noteworthy that only a minority of subjects showed evidence of using model-based RL in session 3, despite extensive experience, a relatively small state space, and fixed transition probabilities. This further suggests that, in domains where subjects do not have strong prior expectations about causal relationships, model-learning is slow and model-free RL dominates adaptive behavior.

Giving subjects explicit information about task structure strongly boosted the influence of model-based RL, as assessed by regression analysis, as well as RL model fitting. This is consistent with meta-cognitive cost-benefit decision making, since giving subjects an accurate model of the task structure will boost the estimated accuracy of model-based predictions and hence the expected payoffs from model-based control. Information about task structure also reduced the influence of, and, unexpectedly, affected model-free control, enhancing the influence on first-step action value updates of the second-step state value relative to the trial outcome, as indexed by the RL model eligibility trace parameter. There is no obvious normative reason why information about task structure should change the use of eligibility traces. Nonetheless, the effect of instruction on the eligibility trace parameter was robust, and was replicated in both clinical groups, as well as healthy volunteers. We suggest that this relationship was not be mediated by changes in a model-free eligibility trace, but rather by changes in representation of the task state-space. Typically, when RL is applied to decision neuroscience, the behavioral task is considered to have a fixed set of discrete states known to the subjects, who are explicitly told the task structure from the start. However, in tasks where subjects must learn task structure from experience, this entails simultaneous and online acquisition of the environments' state-space and of the values of states and actions, from complex and often ambiguous information. The 'model' of the environment that is learned thus comprises not only the state-action-state transition model, used in model-based RL, but also beliefs about the set of states that exist and the current state of the environment, used both by model-based and model-free RL. According to this account, explicit information about reward probability depending on the second-step state will make these states more distinct or salient in internal task representation, such that they are more able to accrue value, which can then drive model-free learning at the first step. This hypothesis should be directly tested in future work. A second unexpected effect of instructions was an increased tendency to repeat choices, as indexed by the RL model perseveration parameter. This likely reflects a strategy of repeatedly sampling a single option to overcome stochasticity within the task. Such sampling may be increased by instruction because subjects have a discrete set of hypotheses that they are deciding between (left is good or right is good), potentially increasing the perceived value of repeated sampling.

Our findings build on extensive literature examining how instruction and experience interact to determine human behavior. Early work examining instruction effects on operant conditioning found that explicit information about the schedule of reinforcement strongly affects responding, such that responses match the contingencies explained to subjects (e.g. fixed interval, variable interval or fixed ratio), even when these differ substantially from the actual contingencies[193–195]. In common with our study, these results emphasize how providing explicit information about task structure allows humans to act in a way that respects that structure, much more readily than via trial and error learning. More recent work has focused on the effect of advice on reward guided decision making in probabilistic settings – i.e. informing subjects that one option is particularly good or bad[198,199]. A central finding from this work is that such advice impacts not only initial estimates of how good or bad different options are, but also modifies subsequent learning by up-weighting and down-weighting outcomes according to the advice given. Whether such bias effects extend to learning about task structure, in addition to simple reward learning, is an open question for further work. Functional neuroimaging has also started to provide mechanistic insight into instruction effects on reward and aversion learning, with findings that instructions change responses to outcomes in the striatum and the ventral prefrontal cortex (ventromedial prefrontal cortex and orbitofrontal cortex), potentially mediated by instructed knowledge represented in the dorsolateral prefrontal cortex[197,200,240]. Our task provides a potential tool for extending such mechanistic investigation of instruction effects into the domain of task structure learning in model-based control.

The results discussed above were those obtained with the Fixed version of the task. Both learning and instruction effects were different in the more complex Changing task, where the transition probabilities linking the first-step actions to second-step states occasionally reversed. In this version there was no increase of model-based control with experience, with model-comparison indicating that mixture and pure model-free models fit the data equally well. The reduced importance of model-based RL in the Changing relative to the Fixed task likely reflects increased uncertainty in the model-based system regarding the transition structure. However, giving subjects explicit information about Changing task structure did not increase use of model-based RL, as indexed by either the regression analysis or the RL model fitting. This suggests that subjects either did not understand the task structure as it was explained to them, or decided that the payoff from model-based control was not worth the effort of tracking both reward and transition probabilities. In fact, loading on the regression analysis outcome predictor, associated with model-free RL, was higher following instruction, suggesting that the main effect from instructions may have been to signal that the environment was volatile.

Unlike the original two-step task, where model-based and model-free RL achieved similar reward rates[238,241], in both task variants used in the current study, use of model-based RL was positively correlated with reward rate. While such correlations between strategy and reward rate can depend on learning rates and choice stochasticity, as shown by Kool and colleagues[241], our results confirm that, for normative human behavior, the relationship can be present. This is desirable in the assessment of this specific task, because it ensures that it reflects the trade-off between accuracy and demand that is an essential part of the balance between model-based and model-free RL systems[241]. Furthermore, in the original two-step task, working memory capacity is correlated with the use of a model-based strategy, which may be related to the capacity to store and retrieve information gathered before the task and not during the task *per se*[187,191]. We did not find any such correlation in either version of the current task (data not shown), possibly due to the fact that, in our study, no instructions were given initially regarding the structure of the task.

Previous evidence collected with the original two-step task indicated that patients with OCD have a deficit in model-based learning[184,242]. However, in our task, although such patients, either medicated or unmedicated, had a deficit in acquiring a model-based strategy directly from task experience, the observed increase in model-based control following debriefing was not different from that found in healthy volunteers. This suggests that patients with OCD do not have a fundamental deficit in the acquisition of model-based control, but rather that they may be unable to understand the instructions for the original two-step task or, with greater likelihood, that in those circumstances OCD patients may have been unable to access and use such information for model-based planning. It also suggests that previous experience in a task environment may facilitate the use of explicit knowledge about task structure. Finally, it is also possible that OCD symptoms do not allow patients to learn a model in some circumstances, but allow them to learn a model in others. Nevertheless, it is important to note that we did find a deficit in model-based learning prior to the delivery of instructions, among patients with OCD. However, this deficit was also present among patients with other mood and anxiety disorders, demonstrating that it is unspecific, possibly related to common mood and anxiety symptoms, or their underlying mechanisms. Our results do suggest, however, a potentially more specific deficit for patients with OCD, that demonstrate a tendency to increase their use of model-free control when learning exclusively from experience, that was not found in healthy volunteers or patients with other psychiatric diagnoses. Further support for enhanced model-free learning in patients with OCD results from the fact that debriefing resulted in a non-significant trend towards a decrease in the model-free parameter, while both healthy volunteers and patients with other mood and anxiety disorders had a clear and significant decrease However, between-group comparisons, while trending towards significance, did not reach significance, suggesting that better powered studies, or studies with adapted task designs, may be necessary to better explore the possibility that increased use of model-free RL, particularly during uninstructed action learning, may be a specific finding in OCD.

In summary, we have developed a new multi-step decision task which allows the effects of learning and explicit information on RL strategy to be dissociated, to our knowledge for the first time. We provide evidence that in this task, use of model-based RL emerges with learning in a subset of individuals, but model-free RL maintains a strong influence on behavior throughout. Information provided to subjects about the task structure increased the use of model-based RL, and reduced the use of model-free RL, but also shaped model-free value updates and exploration/sampling strategies. Finally, we demonstrated the possibility of using this task in clinical populations, and collected data clarifying the RL profile for patients with OCD, with unspecific findings of deficient model-based learning, and more specific findings of enhanced model-free learning, in both cases prior to information about task structure. We suggest that the relationship between use of model-based RL and model-free value updating, observed both in the effect of explicit information and in cross-subject, and possibly cross-disorder, variability in learned behavior, reflects differences in how subjects represent the tasks state-space, specifically whether they treat the two second-step states as real or distinct for the purpose of value learning. The new task's ability to dissociate effects of implicit and explicit information on RL strategy may offer further insight into the content of learning, and provide translational insight on the importance of RL in neuropsychiatric disorders.

3.6. Methods

3.6.1. Participants and testing procedures

The research protocol was conducted in accordance with the declaration of Helsinki for human studies of the World Medical Association and approved by the Ethics Committees of the Champalimaud Centre for the Unknown, NOVA Medical School, Centro Hospitalar Psiquiátrico de Lisboa (CHPL) and New York State Psychiatric Institute. Written informed consent was obtained from all participants. Clinical samples were recruited at the Champalimaud Clinical Centre (CCC), CHPL and the New York State Psychiatric Institute (NYSPI). In each of these centers, patients with OCD were recruited sequentially from clinical or research databases. A mood and anxiety disorders control group was recruited randomly (CCC and CHPL) or sequentially (NYSPI) among patients with the following diagnoses: major depressive episode or disorder, dysthymia, bipolar disorder, generalized anxiety disorder, post-traumatic stress disorder, panic disorder or social anxiety disorder.

Healthy controls were recruited sequentially as a convenience sample of community-dwelling participants, and tested at the same locations

Following consent, each participant was screened for the presence of exclusion criteria, specifically: acute medical illness; active neurological illness; clinically significant focal structural lesion of the central nervous system; history of chronic psychosis, dementia, developmental disorders with low intelligence quotient or any other form of cognitive impairment and illiteracy. Active psychiatric illness, including substance abuse or dependence, was also an exclusion criterion, with the exception of the diagnoses defining inclusion in the OCD and the mood and anxiety groups. In the absence of exclusion criteria, each participant then performed the reduced two-step task (see below). Participants then performed a battery of structured interviews, scales and self-report inventories, including the MINI Neuropsychiatric Interview[213], the Structured Clinical Interview for the DSM-IV[211], the Yale-Brown Obsessive-Compulsive Scale-II (Y-BOCS-II)[243] and the State-Trait Anxiety Inventory (STAI)[217]. In the groups recruited in Lisbon, the Beck Depression Inventory-II (BDI-II) [215,216] was also applied to assess depressive symptoms and the Corsi block-tapping task was used to assess working memory[244], while in New York, the Depression Anxiety Stress Scales (DASS)[245] was applied to assess symptoms of depression, anxiety and stress.

3.6.2. Reduced two-step task

The reduced two-step task was implemented in MATLAB R2014b with Psychtoolbox (Mathworks, Inc., Natick, Massachusetts, USA). The task consisted of a self-paced computer interface with 4 circles always visible on the screen: 2 central circles (upper and lower) flanked by two side circles (left and right) (Fig. 13). Each circle was colored yellow when available for selection, and black when unavailable, and could be selected by pressing the corresponding arrow key (up, down, left or right) on the computer keyboard. Each trial started with both of the central (upper and lower) circles turning yellow, prompting a choice between the two (Fig. 13A). This first step choice then activated one of the side circles in a probabilistic fashion, according to a structure of transition probabilities described below (Fig. 13B). The active side circle could be selected with the corresponding arrow key, resulting either in reward (indicated by the circle changing to the image of a coin) or no reward (indicated by the circle changing to black). The reward probabilities on the right and left side changed in blocks that were either neutral (p=0.4 on each side) or non-neutral (p=0.8 on one side and p=0.2 on the other; Fig. 13C). Changes from non-neutral blocks were triggered based on each subject’s behavior, occurring 20 trials after an exponential moving average (tau = 8 trials) crossed a 75% correct threshold. In half of the cases this led to the other non-neutral block (reward probability reversals), and the other half to a neutral block. Changes from neutral blocks occurred with 10% probability on each trial after the 40th of that block, and always led to the non-neutral block that did not precede that neutral block. All participants performed 1200 trials on the same day, divided in 4 sessions of 300 trials each.

We ran two variants of the task which differed with respect to whether the transition probabilities linking the first-step actions to the second step states were fixed or underwent reversals. In both cases these probabilities were defined such that choosing one of the central circles (e.g. high) would cause one of the side circles (e.g. left) to turn yellow with high probability (p=0.8 – common transition), while causing the other side circle to turn yellow only in a minority of trials, i.e., with low probability (p=0.2 rare transition). Choosing the other central circle would lead to common and rare transitions to the opposite sides. In the Fixed task, the transition probabilities were fixed for each individual throughout the entire task (e.g., common transitions for high-left and low-right, and rare transitions for high-right and low-left). In the Changing task, the transition probabilities underwent reversals on 50% of reward probability block changes after non-neutral blocks, such that the common transition became rare and vice versa (Fig. 13B). In an initial group of healthy volunteers recruited in Lisbon, subjects were randomized between the two versions of the task. In all clinical samples as well as healthy volunteers from New York, however, only the Fixed task was used.

Prior to starting the task, subjects were given minimal information about task structure. They were only told that arrow keys could be used to interact with the screen, and that the image of a coin signaled accrual of a monetary reward. To test how providing explicit information about the task structure affected behavior, debriefing was provided between the 3rd and the 4th sessions in some participants, with the 4th session of the task performed immediately after debriefing. Among healthy volunteers recruited in Lisbon and randomized between the two versions of the task, debriefing was performed in 17 participants performing the Fixed version and in 16 participants performing the Changing version of the task. In all other samples, debriefing was performed for everyone. Please see supplementary material for the specific information provided to subjects prior to the task and during debriefing.

3.6.3. Data analysis

Data analysis was performed using Python (Python Software Foundation, http://python.org) and SPSS (Version 21.0, SPSS Inc., Chicago, IL, USA), and was centered on three main analyses of behavior on the task: reversal analysis, logistic regression analyses of stay probability, and RL model comparison and fitting. The reversal analysis assessed overall task performance according to the average first step choice trajectory around reward probability reversals between non-neutral blocks, from which we extracted two measures. One was the fraction of correct choices at the end of the block before the reversal, with correct defined as the first step choice with a common transition to the side (i.e., state), with highest reward probability. The second measure was the time constant of adaptation to the reversal, estimated by a least squares exponential fit to the cross subject mean choice trajectory following the reversal. For the Changing version of the task, reversal analysis was performed according to the average trajectory for reversals in both transition and reward probabilities. Importantly, while reversal analyses provide information about how well subjects are able to track which option is correct, they do not differentiate between use of model-free and model-based strategies.

The first analysis used to assess model-free vs. model-based behavioral strategies was an analysis of ‘stay-probability’[181,187,188,190–192,229–234], defined as the probability of repeating the first-step choice on any given trial as a function of the outcome (rewarded or not) and transition (common or rare) on the previous trial. In addition to plotting raw stay probabilities, we analyzed the effect of trial events on the subsequent choice using a logistic regression model with several binary predictors. The *Outcome*, *Transition* and *Transition-outcome interaction* predictors modeled the influence of the previous trial’s outcome, transition and their interaction on the probability of repeating the previous first step choice. We additionally included a *Bias* predictor capturing bias towards the upper or lower circle, and a *Correct* predictor, which modeled the influence of whether the previous trials choice was correct (i.e. high reward probability) on the probability of repeating that choice. The latter prevents spurious loading on the *Transition-outcome interaction* predictor, which has been described in two-step tasks with high contrast between good and bad options, due to correlation between action values at the start of the trial and subsequent trial events[238].

Additional analyses of behavioral strategy were obtained by fitting reinforcement learning models to observed behavior. We first detail the model used for the main analyses then a set of alternative models that were rejected by model-comparison. The model followed those typically used in analysis of the original two-step task[181] in combining a model-based and a model-free RL component, both with value estimates contributing to behavior. The model-free component maintained estimates of the values of the first-step ($Q^{mf}(s_1, a_1)$) and second step actions ($Q^{mf}(s_2, a_2)$). These action values were updated as $Q^{mf}_{t+1}(s_1, a_1) = \left(1 - \alpha_Q\right)Q^{mf}_{t}(s_1, a_1) + \alpha_Q((1-\lambda)Q^{mf}_{t+1}(s_1, a_1) + \lambda R)$, and $Q^{mf}_{t+1}(s_1, a_2) = \left(1 - \alpha_Q\right)Q^{mf}_{t+1}(s_1, a_2) + \ \alpha_Q R$, where R is the reward obtained on trial *t* (1 or 0), $\alpha_Q$is the value leaning rate and λ is the eligibility trace parameter. The model-based component maintained estimates of the transition probabilities linking the first step actions to the second step states ($P(s_2|a_1)$), updated as $P_{t+1}(s_2|a_1) = (1-\alpha_T)P_t(s_2|a_1) + \alpha_T$ and $P_{t+1}(s_2'|a_1) = (1-\alpha_T)P_t(s_2|a_1)$, where α_T is a learning rate for transition probabilities, s_2 is the second step state reached and s_2' the second step state not reached on trial *t*. At the start of each trial, model-based action values were calculated as $Q^{mb}_t(s_1, a_i) = \sum_j P\left(s_j \middle| a_i\right) Q_{mf}(s_j, a_2)$. Model-free and model-based action values were combined with perseveration and bias to given net action values, calculated as $Q^{net}_t(s_1, a_i) = G_{mf}Q^{mf}_t(s_1, a_i) + G_{mb}Q^{mb}_t(s_1, a_i) + bB_i + pP_i$, where G_{mf} and G_{mb} are parameters controlling, respectively, the strength of influence of model-free and model-based action values on choice, *b* is a parameter controlling the strength of choice bias, B_i is a variable which takes a value of 1 for the high action and zero for the low action, *p* is a parameter controlling the strength of choice perseveration, P_i is a variable which takes a value of 1 if action a_i was chosen on the previous trial and 0 if it was not. The model's probability of choosing action a_i was given by $P(s_1, a_i) = \frac{e^{Q^{net}(s_1,a_i)}}{\sum_j e^{Q^{net}(s_1,a_j)}}$.

For model comparisons several variants were considered. For the *Model-free only* variant the model-based component was removed such that the net action values were $Q_t^{net}(s_1, a_i) = G_{mf} Q_t^{mf}(s_1, a_i) + bB_i + pP_i$. For the *Model-based only* variant the model-free component was removed such that the net action values were $Q_t^{net}(s_1, a_i) = G_{mb} Q_t^{mb}(s_1, a_i) + bB_i + pP_i$. For the *No bias* variant the bias strength variable b was set to zero. For the *No perseveration* variant the perseveration strength variable p was set to zero.

Fits of both the logistic regression models and the reinforcement learning models to data from populations of subjects used a Bayesian hierarchical modelling framework[246], in which parameter vectors $\boldsymbol{h}_i$ for individual sessions were assumed to be drawn from Gaussian distributions at the population level with means and variance $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. The population level prior distributions were fit to their maximum likelihood estimate $\boldsymbol{\theta}^{ML} = argmax_{\boldsymbol{\theta}}\{p(D|\boldsymbol{\theta}) = argmax_{\boldsymbol{\theta}}\{\prod_i^N \int d\,\boldsymbol{h}_i\, p(D_i|\boldsymbol{h}_i)p(\boldsymbol{h}_i|\boldsymbol{\theta})\}$. Optimization was performed using the Expectation-Maximization algorithm with a Laplace approximation for the E-step at the k-th iteration given by $p\left(\boldsymbol{h}_i^k \middle| D_i\right) = N(\boldsymbol{m}_i^k, \boldsymbol{V}_i^k)$ and $\boldsymbol{m}_i^k = argmax_{\boldsymbol{h}}\{p(D_i|\boldsymbol{h})p(\boldsymbol{h}|\boldsymbol{\theta}^{k-1})\}$, where $N\left(\boldsymbol{m}_i^k, \boldsymbol{V}_i^k\right)$ is a normal distribution with mean $\boldsymbol{m}_i^k$ given by the maximum a posteriori value of the session parameter vector $\boldsymbol{h}_i$, considering the population level means and variance $\boldsymbol{\theta}^{k-1}$, and with covariance $\boldsymbol{V}_i^k$ given by the inverse Hessian of the likelihood around $\boldsymbol{m}_i^k$. For simplicity we assumed that the population level covariance $\boldsymbol{\Sigma}$ had zero off-diagonal terms. For the k-th M-step of the EM algorithm the population level prior distribution parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ are updated as $\boldsymbol{\mu}^k = \frac{1}{N}\sum_{i=1}^N \boldsymbol{m}_i^k$ and $\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^N \left[\left(\boldsymbol{m}_i^k\right)^{\mathbf{2}} + \boldsymbol{V}_i^k\right] - (\boldsymbol{\mu}^k)^2$. Parameters were transformed before inference to enforce constraints $0 < \left\{G_{mf}, G_{mb}\right\}$ and $0 < \left\{\alpha_Q, \alpha_T, \lambda\right\} < 1$. 95% confidence intervals on population means $\boldsymbol{\mu}$ were calculated as $c_i = \pm 1.96\sqrt{-1/H_i}$ where c_i is the confidence interval for parameter i and H_i is the i-th diagonal element of the Hessian at $\boldsymbol{\theta}^{ML}$ with respect to $\boldsymbol{\mu}$.

To compare the goodness of fit for hierarchical models with different numbers of parameters we used the integrated Bayes Information Criterion (iBIC) score. The iBIC score is related to the model log likelihood $p(D|M)$ as $\log p(D|M) = \int d\boldsymbol{\theta}\ p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}|M) \approx -\frac{1}{2} iBIC = \log p(D|\,\boldsymbol{\theta}^{ML}) - \frac{1}{2}|M|\log\ |\mathrm{D}|$, where |M| is the number of fitted parameters of the prior, |D| is the number of data points (total choices made by all subjects) and iBIC is the integrated BIC score. The log data likelihood given maximum likelihood parameters for the prior $\log p(D|\,\boldsymbol{\theta}^{ML})$ is calculated by integrating out the individual session parameters as $\log p(D|\,\boldsymbol{\theta}^{ML}) = \sum_i^N log \int d\boldsymbol{h}\ p(D_i|\boldsymbol{h})p(\boldsymbol{h}|\,\boldsymbol{\theta}^{ML}) \quad \approx \sum_i^N log \frac{1}{K}\sum_{j=1}^K p\big(D_i\big|\boldsymbol{h}^j\big)$, where the integral is approximated as the average over K samples drawn from the prior $p(\boldsymbol{h}|\boldsymbol{\theta}^{ML})$.

Permutation testing was used to assess the effects of learning (sessions 1 vs. 3) and of instruction (sessions 3 vs. 4) on reversal analyses, logistic regression parameters and RL model parameters. To compare a particular parameter (eg, x) between sessions, we calculated the mean value for each session and calculated the difference in means ($\Delta \mathrm{x}_{true}$). We then constructed an ensemble of 5000 permuted datasets in which the assignments of datapoints to each session was randomized. Randomization was performed within subject, to avoid that one subject be represented multiple times within a single session. For each permuted dataset we repeated the analysis, to calculate $\Delta \mathrm{x}_{perm}$. In the limit of many permutations, the distribution of $\Delta \mathrm{x}_{perm}$ is the distribution of $\Delta \mathrm{x}$ under the null hypothesis that there is no difference between the conditions. The two tailed P value for the observed difference is given by $P = 2\min\left(\frac{\mathrm{M}}{N}, 1 - \frac{\mathrm{M}}{N}\right)$, where N is the total number of permutations and M is the number of permutations for which $\Delta \mathrm{x}_{perm} > \Delta \mathrm{x}_{true}$. Permutation tests were also used to compare learning and debriefing effects between different groups (e.g., debriefing vs. no-debriefing groups, clinical groups vs. healthy controls), specifically by tests of interaction between session number and group. We tested the interactions excluding subjects who were identified as model-based in session 3 and then confirmed if the results were maintained when all subjects were included. To assess the significance of the interaction we calculated $\Delta \mathrm{g}_{true}$ = $\Delta \mathrm{x}_{i,j}^A$ – $\Delta \mathrm{x}_{i,j}^B$, where $\Delta \mathrm{x}_{i,j}^A$ is the difference in parameter x between sessions *i* and *j* in group *A*, and $\Delta \mathrm{x}_{i,j}^B$ is the equivalent measure for group B. We then constructed an ensemble of 5000 permuted datasets, as described above, and calculated $\Delta \mathrm{g}_{perm}$ = $\Delta \mathrm{x}_{i,j}^A$ – $\Delta \mathrm{x}_{i,j}^B$ for each permuted dataset, and compared $\Delta \mathrm{g}_{true}$.

Chapter 4. Development of a protocol to collect fMRI data during performance of a new sequential decision task

4.1. Abstract

The use of computational approaches on fMRI studies of action learning has helped to identify brain areas which can be performing RL-like computations in healthy humans. However, some studies suggest that model-based and model-free processes are implemented in the same circuits while other studies suggest that they are implemented in different circuits. Importantly, it has never been tested how the brain represents state space in sequential decision tasks. Also, to our knowledge, although patients with OCD consistently present evidence of dysfunction in corticostriatal circuits, which are relevant for action learning, no study has ever analyzed brain activity in OCD patients while performing an RL-inspired task. In this chapter, we describe the development of an event-based functional magnetic resonance imaging protocol, allowing the capture, to our knowledge for the first time, of brain activity during uninstructed and instructed sequential action choice using the RL framework. This is desirable because, according to the results presented in the previous chapter, uninstructed behavior in the reduced two-step task is predominantly model-free while instructed behavior is predominantly model-based. Data collected to date suggests that our protocol allows to separate neuronal activity associated with each of the three main events of interest in the reduced two-step task. Specifically, choice events are associated with increased BOLD activity in the left precentral gyrus, which corresponds to the primary motor cortex. Reward delivery, on the other hand, is associated with increased BOLD activity in the nucleus accumbens (ventral striatum). We also found that receipt of explicit information about the task contingencies modifies brain activity in prefrontal areas, specifically, increasing BOLD signal in a cluster extending from the paracingulate gyrus into the frontopolar cortex. We also collected data from one OCD patient, but results from event-based GLM analysis did not survive thresholding for multiple comparisons. Data collection for this experiment is ongoing, with future directions including use of multivariate approaches to tackle limitations imposed by the GLM in analyzing ensembles of voxels, and further extending data collection to clinical populations.

4.2. Introduction

Instrumental conditioning experiments have provided clear evidence that separate goal-directed and habitual action controllers exist in the brain and are supported by distinct neural circuits (Fig. 20)[129,132]. Goal-directed behavior can be impaired by lesions in several regions such as the orbitofrontal cortex[247,248], the prelimbic cortex[152,153,156], the dorsomedial striatum[150] and the mediodorsal thalamus[249]. On the other hand, habitual behavior can be impaired by lesions in the dorsolateral striatum[155,250] and in the infralimbic cortex[251]. The only study in humans with focal lesions in these circuits in which an instrumental conditioning task was used, showed that lesions in the ventromedial prefrontal cortex are associated with impairments in goal-directed behavior[252]. In addition to lesion studies, other methods have rendered support to the idea of separate brain circuits for goal-directed and for habitual behavior. Gremel & Costa recorded neural activity in mice performing an instrumental conditioning task which allowed the animals to shift between goal-directed and habitual actions. The authors found that goal-directed actions increase activity in the DMS and in the OFC while habitual actions increase activity in the DLS[253]. In humans, BOLD activity during outcome devaluation is increased in the vmPFC[162] and BOLD activity during contingency degradation is increased in the vmPFC and in the caudate (the human equivalent of the rodent DMS)[163]. On the other hand, extensive training in an instrumental conditioning paradigm has been shown to be associated with increased activity in the putamen (equivalent to the rodent DLS)[164]. However, as mentioned previously, in the Introduction of this dissertation, the capacity of instrumental conditioning adaptations to induce habits in humans has been increasingly questioned, in terms of both its validity and reliability[165].

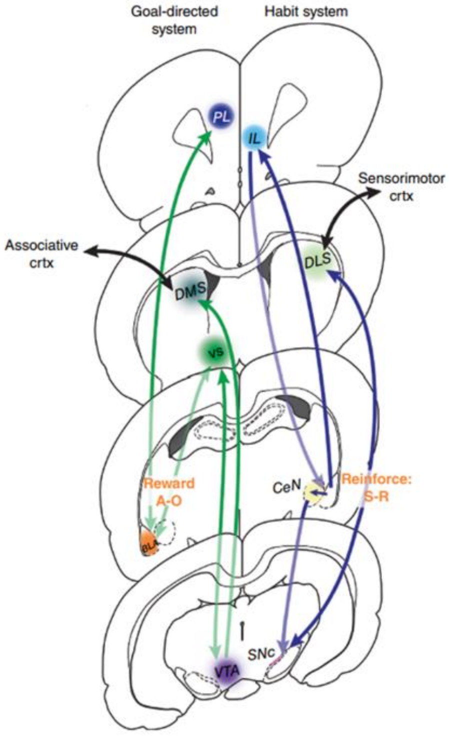


Figure 20. Instrumental conditioning experiments performed in rodents allowed to identify two separate neuroanatomical circuits involved in action selection.

The goal-directed system involves communication between the prelimbic cortex, the dorsomedial striatum, the substancia nigra pars reticulata (SNr) and mediodorsal thalamus, as well as the VTA and the ventral striatum. The habitual system recruits the infralimbic cortex, the sensorimotor cortex, the dorsolateral striatum, the SNr/GPi and the posterior thalamus. From Lingawi et al, 2015[133].

The computational approach provided by RL could offer a solution to that problem. One of the advantages of computational models for functional neuroimaging is that they can be used to define signals of interest[254]. Studies of visual perception, for example, have long benefited from algorithms, such as those derived from signal detection theory, which have helped to make significant advances in experimental design[254]. Although those algorithms have originated in the engineering field, they were used to specify key steps in perception and to search for those computations in the brain. Studies of action control, decision-making and value-based learning took longer to benefit from algorithms which formalize their fundamental steps and specify which quantities to measure and to look for in the brain[254]. Reinforcement learning provides mathematical models of how action control can be implemented. Those models (e. g. model-free control) can be used to generate time series for hidden variables (e. g. reward prediction error) which can be used as regressors (e. g. in a general linear model) to search for voxels which BOLD activity changes accordingly to those regressors[180–183]. The use of computational approaches on fMRI studies of action learning has thus helped to identify brain areas which can be performing RL-like computations[180–183].

However, studies using tasks that allow to differentiate model-based from model-free RL control have shown conflicting results about the neural basis of this systems (Table 6). In the first fMRI application of a decision task based on the RL framework, Gläscher and colleagues found signals associated with a model-free reward prediction error (RPE) in the striatum and signals associated with a model-based state prediction error (SPE) in the lateral prefrontal cortex (and in the intraparietal sulcus)[180]. In contrast, in the first paper using the two-step task, Daw et al. observed signals associated both with model-free and model-based computations in the ventral striatum. In the same participants, BOLD signals associated with both model-free and model-based computations were also found in the prefrontal cortex. Afterwards, Wunderlich and colleagues found signals associated with forward planning – a model-based computation – in the anterior caudate and signals associated with extensive training in the putamen[182]. In contrast, Wimmer et al. observed both model-free and model-based RL correlates in the striatum[255]. These contradictory findings are problematic and are perhaps derived from the fact that these tasks are either good at isolating model-based (e.g. the Glascher task) or model-free (e. g. the Wunderlich task) processes, but are not ideal in having the same subject performing exactly the same task with the same contingencies using model-free or model-based control depending on what he knows about the environment. Our simplified task, which includes a debriefing manipulation, may be able to achieve that because, according to the results presented in the previous chapter, uninstructed behavior in the reduced two-step task is predominantly model-free while instructed behavior is predominantly model-based.

Table 6. Findings from fMRI studies using sequential decision tasks.

Author	Year	Main findings
Integrated (MB and MF in the same areas)		
Daw et al.	2011	MB+MF in lateral PFC and in ventral striatum
Wimmer et al.	2012	MB+MF in ventral striatum and in hippocampus
Separate (MB and MF in the different areas)		
Glascher et al.	2010	MB in lateral PFC and intraparietal sulcus; MF in ventral striatum
Wunderlich et al.	2012	MB in caudate; MF in putamen
Lee et al.	2014	MB in mOFC and ACC; MF in putamen and pre-SMA

MF = model-free BOLD activity. Model-based BOLD activity. PFC = prefrontal cortex. vmPFC = ventromedial prefrontal cortex. mOFC = medial orbitofrontal cortex. ACC = anterior cingulate cortex. Pre-SMA = pre-supplementary motor area.

Also, the impact of instructions on how subjects represent sequential decision tasks in the brain is not known. It has been shown, using a probabilistic learning task, that instructed knowledge of cue-reward probabilities improves performance and decreases BOLD responses in the ventral striatum, ventromedial PFC and hippocampus[200]. The decrease in activity in these regions is correlated with activity in the dorsolateral PFC, leading authors to suggest that humans use the DLPFC to dynamically adjust outcome responses in valuation regions depending on the usefulness of the action-outcome information. The impact of instructions on aversive learning has also been studied using fMRI[197]. Atlas and colleagues have found that telling participants about reversals in the contingencies between image presentations and mild electric shocks caused changes in the activity of the striatum and the prefrontal cortex[197]. However, we are not aware of fMRI studies which use sequential decision tasks to explore the impact of explicit knowledge on RL action control systems. In fact, one of our unexpected findings with the behavioral version of reduced two-step task also raises an important question which could be answered using neuroimaging: the debriefing led to a consistent decrease in the eligibility trace parameter across all groups, suggesting that explicit information about task structure was not only used by the model-based system but also by another system in the brain which should be representing the state space of the environment. Also, no study has ever used fMRI during performance of an RL-inspired sequential decision task in OCD patients. The study which applied the original two-step task in OCD patients did not collect imaging data in the OCD group[235]. Nevertheless, structural MRI data was collected in healthy volunteers and a positive correlation between use of model-based control and left OFC volume was found[235]. This is particularly interesting because, as reviewed in the Introduction of this thesis, the OFC is the brain area which has most consistently been shown to be dysfunctional in OCD, with several studies and meta-analysis showing hyperactivity at rest and reduced structural volume[61].

However, brain circuitry dysfunction in OCD is far more complex than orbitofrontal cortex (OFC) hyperactivity. First of all, because the OFC is not homogenous in the healthy brain, neither anatomically nor functionally, with medial and lateral sectors with different projections and functions[256–259]. Second, because the lateral OFC shows increased activity during classical symptom provocation studies[260] but reduced activity during reversal learning tasks[261], in which OCD patients have shown deficits. Third, because it has been demonstrated that the connectivity between the prefrontal and striatal areas is altered at rest, with a pattern of increased functional connectivity between the OFC and the nucleus accumbens (ventral striatum) and reduced functional connectivity between lateral PFC and the caudate (dorsomedial striatum) identified using resting-state fMRI[262]. Fourth, and extremely relevant for the context of this dissertation, because a recent symptom-provocation fMRI study inspired by the habit hypothesis of OCD showed an interesting pattern of activity in the OFC and in striatum during the moment when OCD patients performed an action in order to stop OC-inducing stimuli[263]. Banca and colleagues measured, on a trial by trial basis, patient's responses to the visualization or contact with stimuli that elicited OC symptoms. Crucially, in a very elegant behavioral manipulation, participants could choose to terminate the presentation of the symptom-provoking stimuli by performing an action – pressing a button. According to the authors, this action could be seen as an analogous to a compulsion (which reduces the anxiety caused by an obsession) and should be associated with activity in dorsolateral striatal (putamen) areas. Their results show that symptom-provoking conditions evoked a dichotomous pattern of deactivation/activation in OCD patients, which was not observed neither in control conditions nor in healthy subjects: a deactivation of caudate-medial OFC circuits (associated to goal-directed actions) accompanied by hyperactivation of subthalamic nucleus/putaminal regions (associated to habitual actions)[263]. Importantly, the putaminal hyperactivity during patients' symptom provocation preceded subsequent deactivation during the act of responding to end the symptom-provoking condition. Although these results are extremely interesting, participants were not facing a sequential decision problem. Thus, it was not possible to further analyze this data, through a computational perspective, which could allow to search for BOLD signals that changed according to the predictions of a RL algorithm.

4.3. Objectives

4.3.1. Develop a protocol to collect fMRI data during performance of the reduced two-step task

4.3.2. Investigate if, in the fMRI version of the reduced two-step task, pre-debriefing performance reflects a higher reliance on model-free control and post-debriefing performance reflects a higher reliance on model-based control

4.3.3. Explore differences between brain activity when learning exclusively from experience and brain activity during instructed performance of the same task in the same subject

4.4. Methods

4.4.1. fMRI data collection

Imaging data was acquired on a 3 T SIEMENS MAGNETOM Prisma scanner. Functional data was acquired using a T2*-weighted echo planar imaging sequence [repetition time (TR) = 2 s, echo time (TE) = 30 ms, flip angle (FA) = 75°, field of view (FOV) = 212 mm]. Seventy-two oblique axial slices were acquired with a 2 mm in-plane resolution positioned along the anterior commissure-posterior commissure line. Slices were acquired in an interleaved fashion. Each run consisted of 25 trials of the reduced two-step task. In addition to functional data, a single three-dimensional high-resolution (1 mm isotropic) T1-weighted full-brain image was acquired using a MP RAGE pulse sequence for brain masking and image registration.

4.4.2. Reduced two-step task adaptation for fMRI

The structure of the reduced two-step task was modified to implement a task design with intra-trial and inter-trial intervals adapted to event-based fMRI data collection. The most important modifications were made in order to resolve BOLD signals at three specific time points in each trial: 1) during first-step choice ("choice" event), 2) when the second step is revealed ("transition" event), 3) when the outcome is revealed ("outcome" event). To increase our ability to detect signals associated with these three events of interest, we added intermediate events between them (Fig. 21). After the first-step choice was made, the circle which the participant chose was highlighted with a white circular border around it. After the second-step action was made, the corresponding circle was similarly highlighted. In order to sample the hemodynamic response function at different timepoints in each trial, the duration of these intermediate events was jittered, rather than constant, varying between 2 and 8 seconds, and with a mean duration of 3 seconds. To further increase the accuracy of the estimation of the hemodynamic response curve, we also implemented jittered inter-trial intervals (mean=3s, min.=2s, max=8s). Importantly, even though the duration of the intermediate events was independent of the participants' actions, the fMRI version of the task is self-paced at both the first- and second-steps.

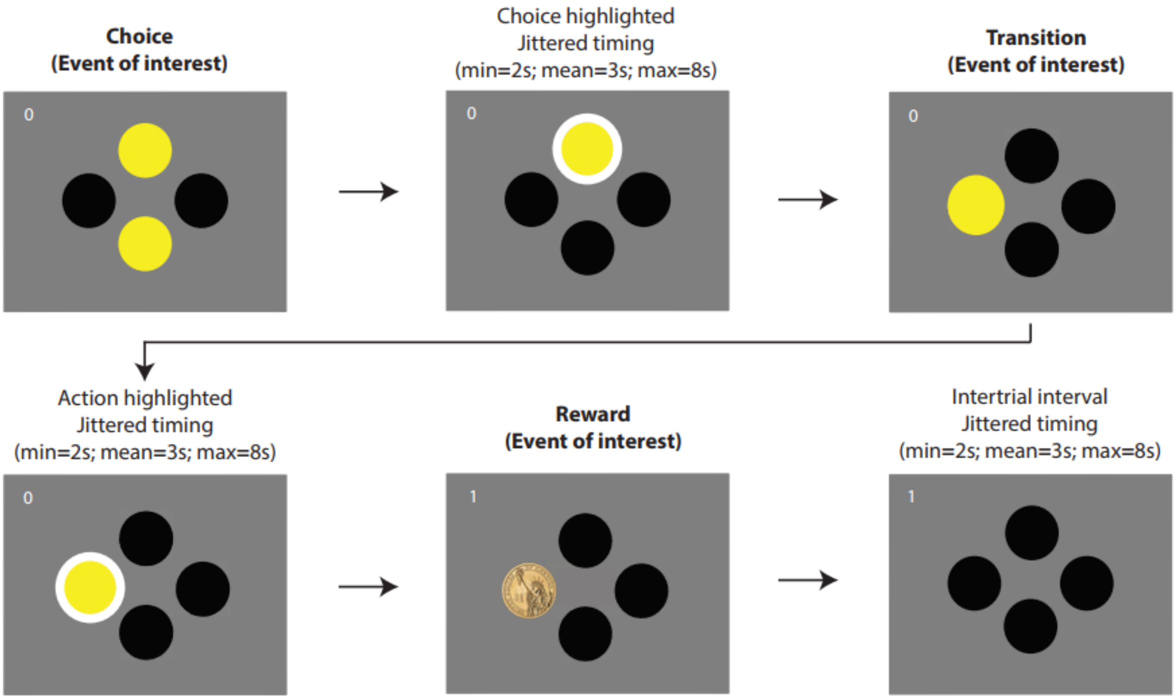


Figure 21. fMRI adaptation of the reduced two-step task – trial events.

The main task structure is maintained (see Fig. 14 – Reduced two-step task structure). In order to resolve BOLD activity in each event of interest (first-step choice; transition; reward delivery), intermediate events were added between those events. After the participant has chosen between the upper circle and the lower circle at the first-step, his choice is highlighted with a white circle around it. After the participant has pressed the button corresponding to the second-step state reached (left or right), the corresponding circle is highlighted with a white circle around it. The duration of these intermediate events follows a jittered timing with a mean of 3 seconds (minimum 2 seconds, maximum 8 seconds). The duration of the intertrial intervals also followed a jittered timing.

The task modifications implemented to deal with the fMRI constraints increased the duration of each trial, and thus the duration of the task, during which participants should remain within the scanner. We thus decided to decrease the number of trials to 200 before and 100 trials after the debriefing. Pilot tests of versions with a higher number of trials resulted in a very long duration of each experiment, with over 2 hours inside the scanner. In all cases, we preserved the balance between pre-debriefing and post-debriefing trials that we used in the behavioral version described in Chapter 3. The experiment was divided in runs/sessions with a duration of 6.7 minutes on average (Fig. 22), and within each run/session, the participant performed 25 trials. Between runs, participants could rest and relax, and the experimenter can communicate with the participant. No image is collected between runs. One additional modification for the fMRI experiment was to use a fixed pattern of block transitions (i.e. make them independent of the subjects behavior and consistent across subjects), to reduce this source of variability and to ensure that there was good coverage of the different block types over the limited number of trials (Fig. 23).

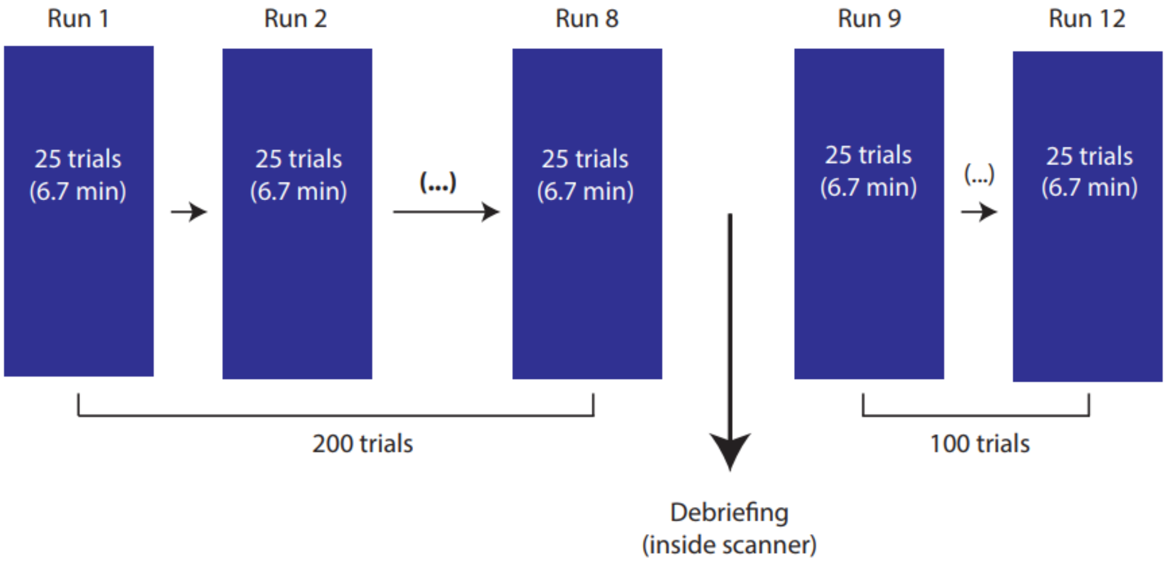


Figure 22. fMRI adaptation of the reduced two-step task – experimental design.

Participants performed 12 runs/sessions of the task consisting of 25 trials in each run/session. Between the eighth and the ninth run, subjects were given explicit information about the task structure (transition probabilities, reward probabilities) in a debriefing presented while inside the scanner.

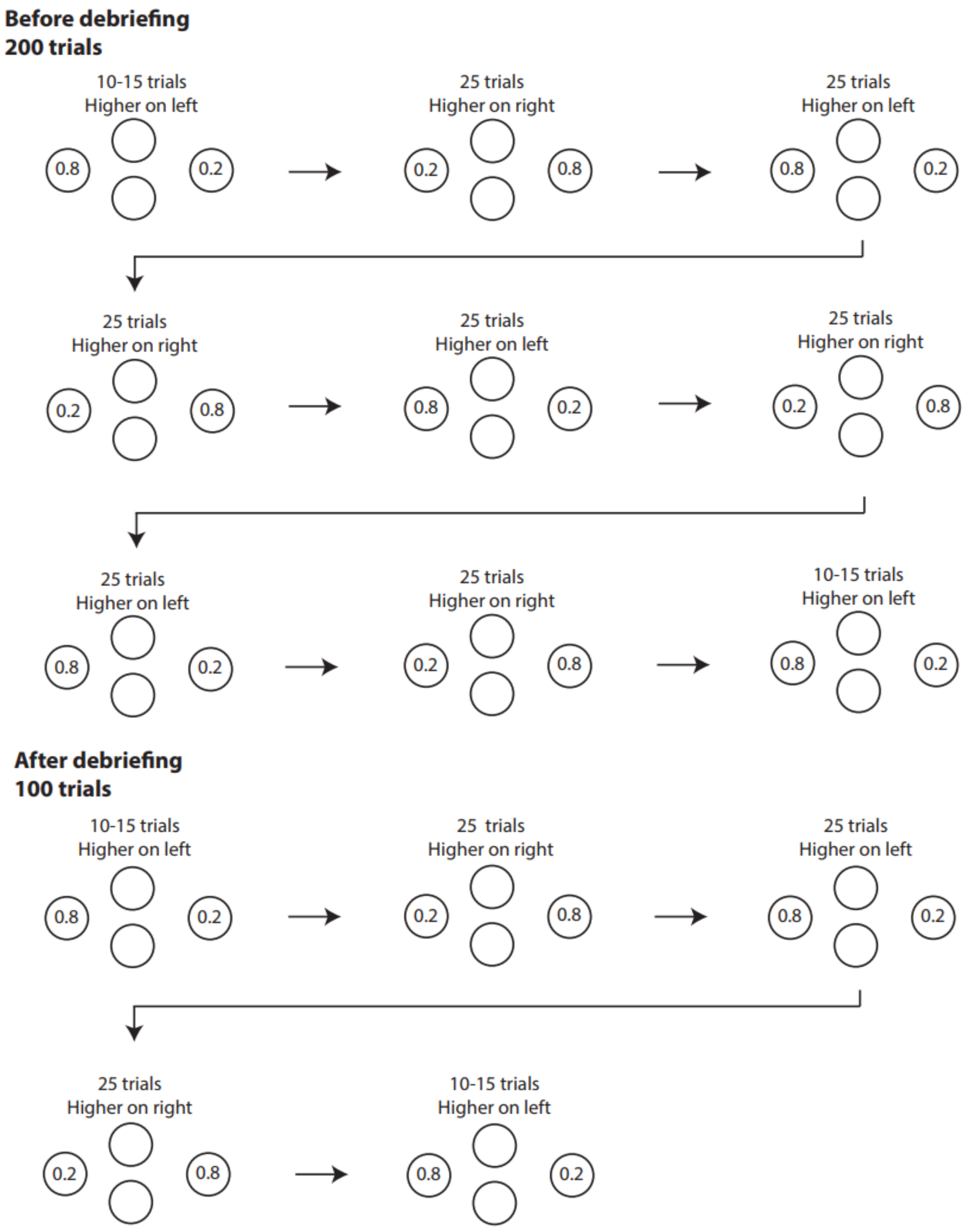


Figure 23. fMRI adaptation of the reduced two-step task – block structure.

We modified the block structure of reward probabilities (which was dependent on performance in the behavioral version presented on Chapter 3 – see section 3.4 Methods) in order to guarantee that all participants experienced one modification of reward probabilities per run/session.

4.4.3. Data analysis

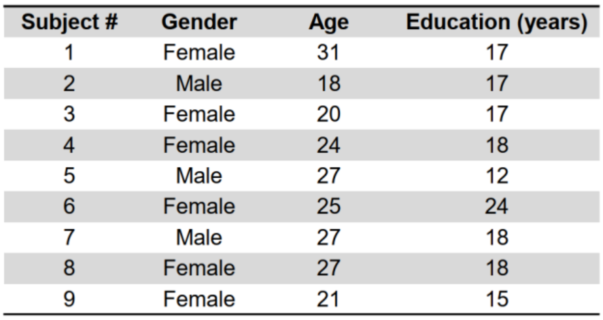
We analyzed behavior using a model-agnostic stay-probability analysis and logistic regression similar to the ones described in section 3.4. To explore the effects of experience in the task, we compared the early runs/sessions (1 to 4) with late runs/sessions (5 to 8) using permutation tests. To explore the effects of explicit knowledge, we compared the late runs/sessions (5 to 8) with the post-debriefing runs/sessions (9 to 12) using permutation tests.

Functional MRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). Registration of the functional data to the high resolution structural image was carried out using boundary based registration algorithm[264]. Registration of the high resolution structural to standard space images was carried out using FLIRT[265,266]. The following pre-statistics processing was applied; motion correction using MCFLIRT[266]; non-brain removal using BET[267]; spatial smoothing using a Gaussian kernel of FWHM 5mm; grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50.0s). Time-series statistical analysis was carried out using FILM with local autocorrelation correction[268]. The time series model included three general linear model (GLM) regressors based on trial events in the task: 1) choice; 2) transition and 3) reward. The “choice” regressor (parameter 1) was defined as the time, in each trial, at which the participant chose between the upper circle or the lower circle at the first-step step of the task and its duration was defined by how long this choice was highlighted (this timing was jittered with a mean=3s; min=2s; max=8s). The “transition” regressor (parameter 2) was defined as the time, in each trial, at which the transition which occurred (common vs. rare) was revealed and its duration was defined by how long the second-step choice was highlighted (this timing was jittered with a mean=3s; min=2s; max=8s). The “reward” regressor (parameter 3) was defined as the time, in each trial, at which the reward delivery (or non-delivery) occurred (at the end of the second step); its duration was defined by how long the reward (coin) was shown on the screen (fixed 1.5s) and a parametric modulation was used by mean-centering the number of rewards obtained across the 25 trials. All regressors were entered at the first level of analysis and all were convolved with a canonical double-gamma hemodynamic response function. The temporal derivative of each regressor as included in the model. The models were estimated separately for each participant and each run. Second-level analysis, combining runs/sessions within subject, was carried out using a fixed effects model, by forcing the random effects variance to zero in FLAME (FMRIB's Local Analysis of Mixed Effects)[269–271]. Correction for multiple comparisons was performed using cluster-based thresholding.

4.5. Results

Recruitment for this experiment is currently ongoing. The results described here are preliminary, and essentially a proof-of-principle regarding the possibility of collecting event-related BOLD-signal in the MRI scanner, during performance of the reduced two-step task. Currently, I have collected fMRI data from 9 healthy subjects, 3 of whom are men, with a mean age of 24 years and a mean education of 17 years (Table 7).

Table 7. Sociodemographic data for participants who performed the fMRI version of the reduced two-step task.



Subject #	Gender	Age	Education (years)
1	Female	31	17
2	Male	18	17
3	Female	20	17
4	Female	24	18
5	Male	27	12
6	Female	25	24
7	Male	27	18
8	Female	27	18
9	Female	21	15

Comparing early pre-debriefing sessions with late pre-debriefing sessions, we observed no significant increases in the ‘outcome’, ‘transition’ or ‘transition x outcome’ logistic regression predictors (Fig. 24). There was a significant increase in the ‘choice’ predictor from early to late sessions, reflecting a higher tendency to repeat the same first-step choice, independently of the trial events. Comparing late pre-debriefing sessions with post-debriefing sessions, we observed a decrease in the ‘outcome’ logistic regression predictor and an increase in the ‘transition x outcome’ predictor (Fig. 25).

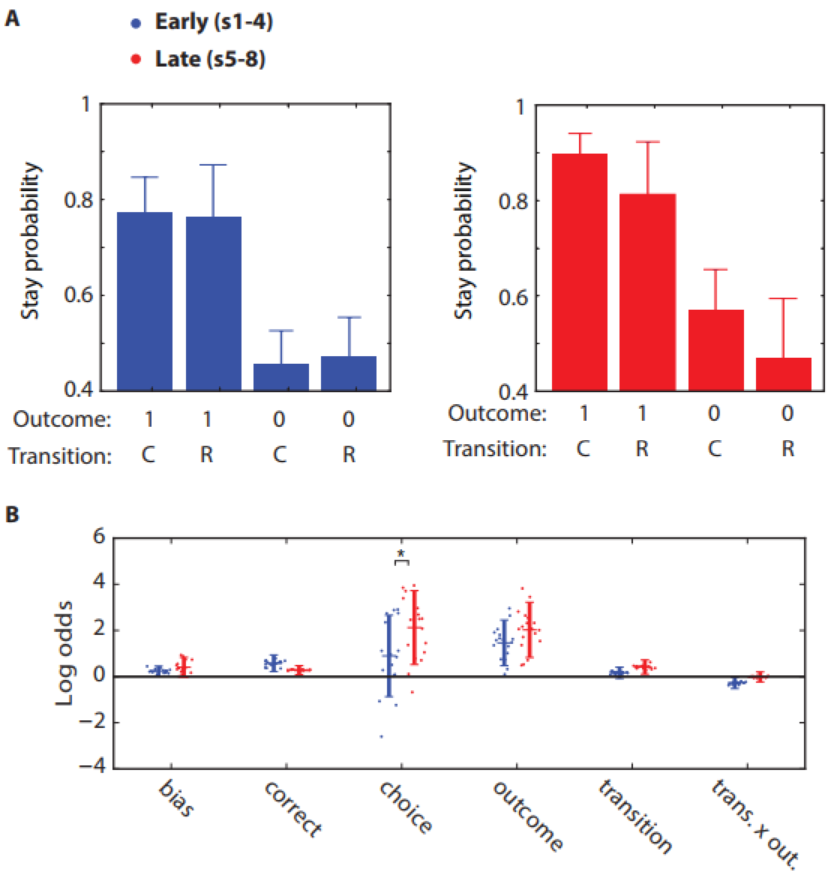


Figure 24. Effects of experience in the fMRI version of the reduced two-step task in healthy subjects

A) Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicate the cross subject standard error (SEM). The top left panel shows data combining early pre-debriefing sessions (session 1 to session 4), the top right panel shows data combining late pre-debriefing sessions (session 5 to session 8). **B)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Positive loading on the ‘outcome’ predictor indicates a tendency to repeat rewarded choices. Positive loading on the ‘transition’ predictor reflects a tendency to repeat choices followed by common transitions. Positive loading on the ‘transition x outcome’ interaction predictor indicates a tendency to repeat choices that were rewarded following a common transition, or that were not rewarded following a rare transition. Additional predictors include ‘bias’, indicating a tendency to choose the upper circle, ‘correct’, which models the influence of whether the previous trials choice was correct (i.e. high reward probability) on the probability of repeating that choice, and ‘choice’, indicating a tendency to repeat the previous first-step choice independently from the previous trial events. Dots indicate maximum a posteriori parameter values for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. Statistical significance of differences in factor loadings for each predictor between early sessions (blue) and late sessions (red) were evaluated using permutation tests.

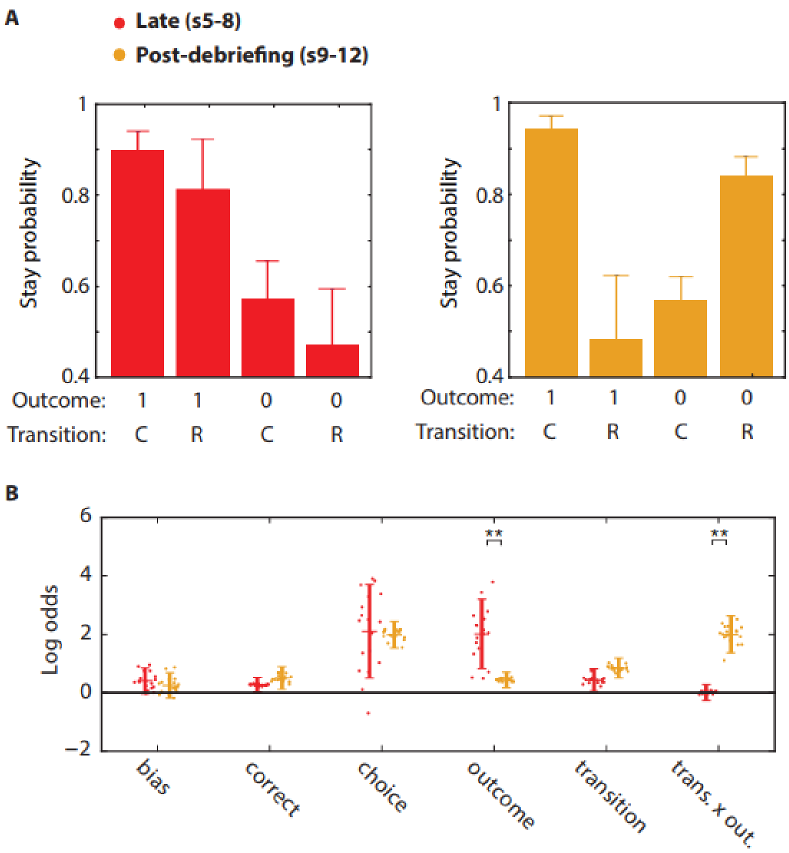


Figure 25. Comparison of late pre-debriefing sessions with post-debriefing sessions in the fMRI version of the reduced two-step task in healthy subjects

A) Stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicate the cross subject standard error (SEM). The top left panel shows data combining late pre-debriefing sessions (session 5 to session 8), the top right panel shows data combining late pre-debriefing sessions (session 8 to session 12). B) Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Positive loading on the ‘outcome’ predictor indicates a tendency to repeat rewarded choices. Positive loading on the ‘transition’ predictor reflects a tendency to repeat choices followed by common transitions. Positive loading on the ‘transition x outcome’ interaction predictor indicates a tendency to repeat choices that were rewarded following a common transition, or that were not rewarded following a rare transition. Additional predictors include ‘bias’, indicating a tendency to choose the upper circle, ‘correct’, which models the influence of whether the previous trials choice was correct (i.e. high reward probability) on the probability of repeating that choice, and ‘choice’, indicating a tendency to repeat the previous first-step choice independently from the previous trial events. Dots indicate maximum a posteriori parameter values for individual subjects, bars indicate the population mean and 95% confidence interval on the mean. Statistical significance of differences in factor loadings for each predictor between late pre-debriefing sessions (red) and post-debriefing sessions (gold) were evaluated using permutation tests.

Regarding the fMRI data, after running the first-level GLM (producing parameter estimates for each of our three events of interest in each session), we combined all 12 runs in a higher-level GLM. Then, to test for the BOLD effects of uninstructed experience in the task, we compared early pre-debriefing sessions (sessions 1 to 4) with late pre-debriefing sessions (sessions 5 to 8). To explore the effects of explicit knowledge on BOLD activity, we compared pre-debriefing runs (run 5 to run 8) with post-debriefing runs (run 9 to run 12).

The preliminary analyses of imaging data show that, across all 12 runs, choice events were associated with increased BOLD activity in the left precentral gyrus (Fig. 26, Table 8), while reward events were associated with increased BOLD activity in the left ventral striatum (nucleus accumbens; Fig. 27). Reward events are also associated with increased activity bilaterally in the fusiform cortex, predominantly on the right side (Table 9). However, no area survived multiple comparisons regarding focal activity associated with transition events (data not shown).

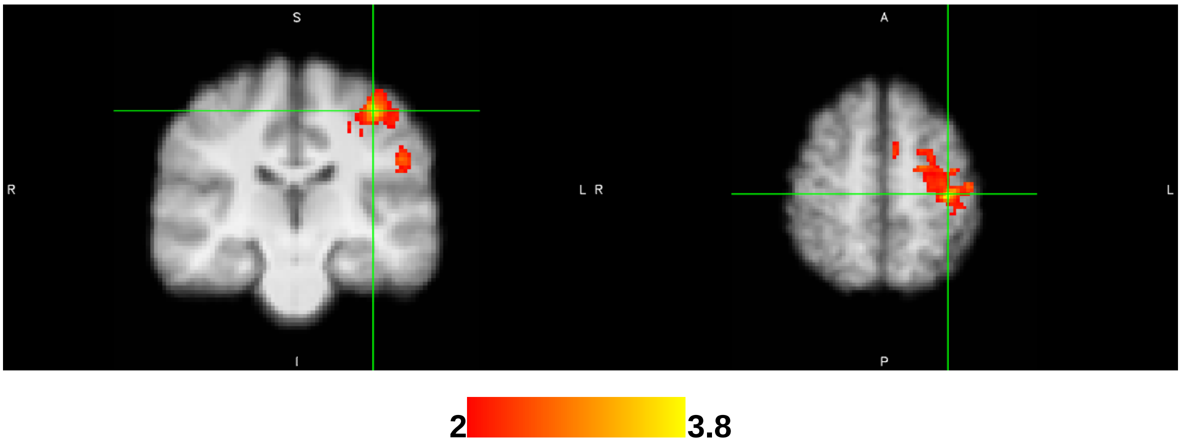


Figure 26. Statistical map showing activation in the left precentral gyrus cortex for choice events in healthy subjects.

Heatmap color bars range from z-stat = 2 to 3.8. Correction for multiple comparisons performed using cluster-based thresholding with clusters determined by z>2.0 and a corrected cluster significance threshold of p<0.05. The maximum intensity voxel in the largest cluster of activation corresponds to the left precentral gyrus (Harvard-Oxford Cortical Structure Atlas).

Table 8. Activation table for map in figure 26.

Cluster Index	Voxels	P	Z-MAX	Z-MAX X (mm)	Z-MAX Y (mm)	Z-MAX Z (mm)
1	1015	0.00057	3.76	-38	-22	56

Cluster Index: a unique number for each cluster from 1 to N. Voxels: the number of voxels in the cluster. P: P-value for each cluster. Z-MAX: the value of the maximum "intensity" within the cluster (the maximum z-statistic). Z-MAX X/Y/Z (mm): the location of the maximum intensity voxel, given as X/Y/Z coordinate values in standard space coordinates.

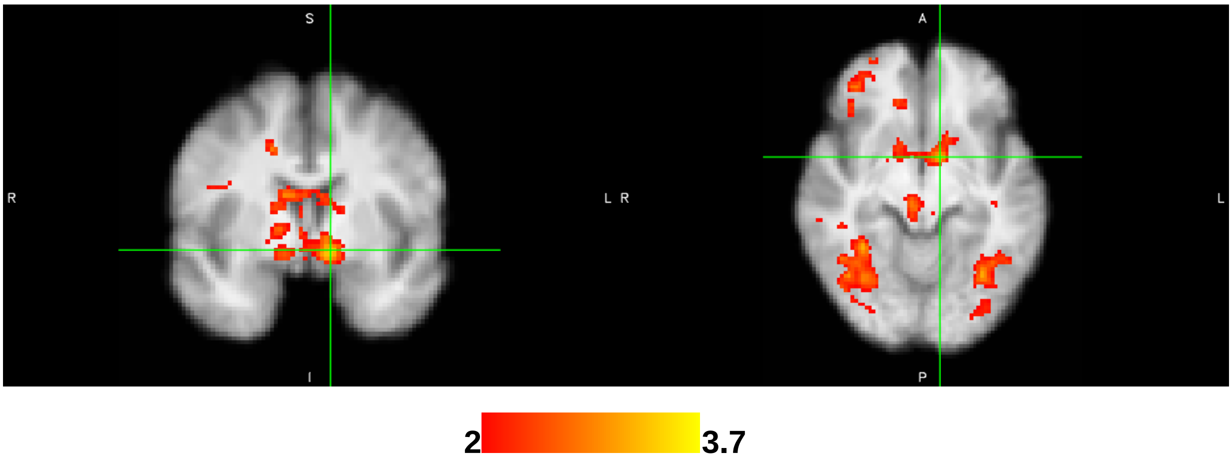


Figure 27. Statistical map showing activation in the left ventral striatum (nucleus accumbens) for reward events in healthy subjects.

Heatmap color bars range from z-stat = 2 to 3.7. Correction for multiple comparisons performed using cluster-based thresholding with clusters determined by z>2.0 and a corrected cluster significance threshold of p<0.05. The maximum intensity voxel in the largest cluster of activation corresponds to the left ventral striatum (nucleus accumbens) (Harvard-Oxford Cortical Structure Atlas).

Table 9. Activation table for map in figure 27.

Cluster Index	Voxels	P	Z-MAX	Z-MAX X (mm)	Z-MAX Y (mm)	Z-MAX Z (mm)
3	5052	1.92E-11	3.48	-10	4	-8
2	1788	4.29E-05	3.68	34	-50	-12
1	1235	0.00105	3.11	-34	-64	-8

Cluster Index: a unique number for each cluster from 1 to N. Voxels: the number of voxels in the cluster. P: P-value for each cluster. Z-MAX: the value of the maximum "intensity" within the cluster (the maximum z-statistic). Z-MAX X/Y/Z (mm): the location of the maximum intensity voxel, given as X/Y/Z coordinate values in standard space coordinates. The largest cluster (5052 voxels) has its maximum intensity in the ventral striatum (nucleus accumbens). The other two clusters (1788 voxels and 1235 voxels) correspond, respectively, to the right and left fusiform gyrus.

Comparing early pre-debriefing sessions (1 to 4) with late pre-debriefing sessions (5 to 8) we found no clusters of activity which survived the multiple comparisons correction (data not shown). Comparing pre- with post-debriefing sessions, we found increased BOLD activity for transition events in a cluster extending from the right paracingulate cortex into the right frontopolar cortex (Fig. 28, Table 10).

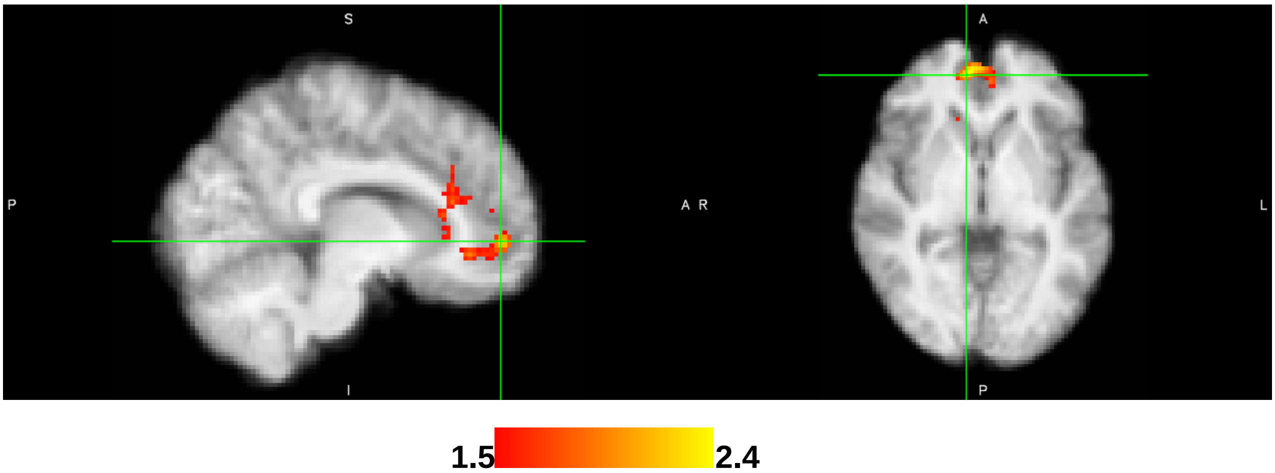


Figure 28. Statistical map showing debriefing effects on transition events in healthy subjects.

Heatmap color bars range from z-stat = 1.5 to 2.4. Correction for multiple comparisons performed using cluster-based thresholding with clusters determined by z>1.5 and a corrected cluster significance threshold of p<0.05. The maximum intensity voxel in the largest cluster of activation corresponds to the right paracingulate cortex (Harvard-Oxford Cortical Structure Atlas).

Table 10. Activation table for map in figure 28.

Cluster Index	Voxels	P	Z-MAX	Z-MAX X (mm)	Z-MAX Y (mm)	Z-MAX Z (mm)
1	1302	0.0474	2.39	8	54	0

Cluster Index: a unique number for each cluster from 1 to N. Voxels: the number of voxels in the cluster. P: P-value for each cluster. Z-MAX: the value of the maximum "intensity" within the cluster (the maximum z-statistic). Z-MAX X/Y/Z (mm): the location of the maximum intensity voxel, given as X/Y/Z coordinate values in standard space coordinates.

4.6. Discussion

Our preliminary results suggest the modifications we implemented in the reduced two-step task were successful in maintaining a behavioral performance which was similar to the version of the task we presented in Chapter 3. The most important effect was the dramatic change in the ‘transition x outcome’ and in the ‘outcome’ predictors of the logistic regression, which suggest that explicit knowledge led to an increase in model-based control and to a decrease in model-free control. Although we found no differences in the main logistic regression predictors when comparing early pre-debriefing sessions with late pre-debriefing sessions, we must keep in mind that we’re only comparing the first 100 trials with the following 100 trials – less than the

300 trials that subjects performed only in the first session of the behavioral version of the reduced two-step task (Chapter 3).

We also demonstrate that the fMRI protocol we developed allows to separate the three main events of interest in our sequential decision task: the first-step choice; the transition from the first- to the second-step and the reward delivery. Moreover, our results regarding modulation of activity during these events is consistent with previous literature. The finding of increased BOLD activity in the left precentral gyrus during choice events is consistent with the well-established role of this area in controlling motor output[272,273]. This finding is reassuring regarding our chosen methods, since “choice” was the only time series which was aligned with an action, and it was associated with increased BOLD activity in the contralateral motor cortex. The increase in ventral striatal (nucleus accumbens) activity during reward delivery is also in agreement with our predictions, given previous literature demonstrating associations between nucleus accumbens activity and reward delivery. Specifically, it has been previously shown that the nucleus accumbens increases its activity in anticipation of reward[274] and, in the original two-step task, Daw and colleagues found BOLD signals in the ventral striatum correlated both with a model-free reward prediction error and a model-based state prediction error[181].

An unexpected finding was the increased neuronal activity in the fusiform cortex, bilaterally, associated with reward events. The fusiform cortex is known for its role in face and body recognition[275–277]. Our interpretation is that reward events increased BOLD activity in the fusiform area because the reward symbol shown was a United States dollar coin, which features an image of the Statue of Liberty (including its face). In support of this hypothesis, the right cluster of activation is larger and more significant than that on the left, just as the fusiform face area is larger on the right than on the left[275–277].

The preliminary data collected until now also suggests that the debriefing may modify brain activity while performing the same task. When comparing pre- with post-debriefing runs, the only significant cluster of activation extended from the paracingulate cortex (part of the medial prefrontal cortex) into the frontal pole. Interestingly, both the medial prefrontal cortex and the frontopolar cortex have been associated with the arbitration process between model-based and model-free RL using sequential decision tasks[182,183]. We will further explore the computations underlying this arbitration process and how debriefing may modify them in Chapter 5.

Due to low number of subjects until the present moment, behavioral data has not yet been incorporated in the GLM analysis. We plan to use a GLM analysis with RL-driven predictors in order to identify brain areas associated with model-based and model-free computations. Our experimental design has advantages compared to others, namely because it should separate runs/sessions in which behavior should be under model-free control (pre-debriefing) from sessions in which behavior should reflect a hybrid of model-based and model-free control (post-debriefing). However, the GLM is a massive univariate approach, posing some limitations. Specifically, it only allows to examine individual voxels or regions at a centimeter scale (the "region of interest" approach). As a result, it is unclear which voxel(s) corresponds to which unit(s) of the computational model, since a one-to-one mapping cannot be assumed because one voxel corresponds to a large number of neurons.

Multivariate approaches, such as multivoxel pattern analysis (MVPA), try to solve the spatial correspondence problem mentioned above by considering spatial patterns of activity over ensembles of voxels to analyze what information they represent collectively[278]. Here I plan to use a specific class of MVPA which focuses on the similarity of voxel patterns, typically called "similarity-based MVPA"[279]. These approaches are not limited to showing that a brain region is *involved* in some function, but rather uses representational models that specify *how* different perceptions, cognitions or actions are encoded in brain-activity patterns[280]. Representation-similarity analysis (RSA) is probably the most versatile version of similarity-based MVPA. It extends beyond analyzing information in regional response patterns and allows testing of conceptual and computational models in ensembles of voxels. RSA characterizes the representation in each brain region by a representational dissimilarity matrix, i.e., a square symmetric matrix with each entry referring to the dissimilarity between the activity patterns associated with two stimuli or experimental conditions, as measured by the Euclidean distance, Pearson's correlation distance, or other method. Representational models are then assessed by comparing the predicted to the observed dissimilarities. We plan to use RSA to analyze how the brain represents task space, contingencies and reward information. Importantly, we plan to test if debriefing changes the pattern of activity in ensembles of voxels located in prefrontal and/or in the striatal areas.

In summary, we have designed an fMRI protocol with a sequential decision task that allows, to our knowledge for the first time, to capture differences in brain activity between sessions in which behavior is predominantly model-free (before instructions) and sessions in which behavior is predominantly model-based (after instructions), in the same subject, task structure and environment. Future directions include expanding the ongoing data collection, use multivariate analysis methods and recruit a sample of OCD patients and controls.

Chapter 5. Discussion and conclusions

In this dissertation I have examined OCD through different complementary approaches. After reviewing the literature in the Introduction, I followed a classical psychometrics framework to establish the criterion validity of the gold-standard instrument for assessing OCD and OCD symptom severity. However, I also confirmed previously reported limitations of this instrument with respect to discriminant validity. I then reviewed the current state-of-the-art with respect to the use of reinforcement learning paradigms to study learning and decision processes in OCD patients, and used computational modeling of behavior to develop a new sequential decision task – the reduced two-step task – which tries to circumvent the main limitations of previously used paradigms. I demonstrated that OCD patients have trouble in increasing their use of model-based RL when learning exclusively from experience in the reduced two step task but are able to use model-based RL control when given explicit information about the structure of the environment. While this same pattern of model-based control deployment was also present in a group of patients with mood and anxiety disorders, an increase in model-free RL during uninstructed performance was only present in OCD patients, which were also the only group that did not decrease their use of model-free RL after obtaining explicit knowledge about the structure of the task. Finally, I developed a protocol to collect functional MRI data during performance of the the reduced two-step task, and demonstrated that this protocol has advantages over other RL-inspired studies by allowing us to compare brain activity in uninstructed sequential action choice with brain activity in instructed sequential action choice. Importantly, I demonstrated that the effects of explicit knowledge on RL action control are also present in the modified version that I developed for use with fMRI. Regarding brain activity, I found that, across all sessions, choice events were associated with increased BOLD activity in the left precentral gyrus, while reward events were associated with increased BOLD activity in the left ventral striatum. Comparing pre- with post-debriefing sessions, I found increased BOLD activity for transition events in a cluster extending from the paracingulate cortex into the frontal pole. I will now discuss each of these main findings, its limitations and several promising future directions.

5.1 The Yale-Brown Obsessive-Compulsive Scale Second Edition has good criterion-validity but moderate divergent validity

Adequate assessment of obsessive-compulsive symptoms is required to attain the main goal for OCD patients and treatment providers: optimal therapeutic outcome[24,25]. Access to the best treatment interventions is based on accurately establishing that a person has OCD. Also, sound psychometric tests are necessary to measure OCD symptom severity if we want to test the efficacy of therapeutic interventions with precision. Two types of psychometric instruments can be used to assess OCD: clinician-rated or patient-report measures [24]. Tests administered by a clinician have the advantage of providing detailed and useful information about the nature and severity of OC symptoms and guarantee that the patient can elaborate on his symptoms or clarify any items. However, as disadvantages, these types of measure need extensive training and take substantial time to implement in research and clinical practice[24]. Several self-administered measures of OC symptom severity are also available. While such measures may have the advantages of practicality, brevity and minimal patient burden, these are widely supplanted by various serious limitations: lack of independent verification of responses, potential for response bias, lack of translation into certain languages or applicability to those with low reading level[24,25]. Also, in patients who have few but very specific obsessions or compulsions, self-report measures may underestimate global severity[25]. In addition to the Y-BOCS/Y-BOCS-II, the National Institute of Mental Health Global Obsessive Compulsive Scale (NIMH-GOCS)[281] and Clinical Global Impressions (CGI)–Severity Scale[282] have been used as clinician ratings of OCD symptom severity and overall illness severity. While advantages of each include the brevity, wide use, and moderate to good psychometric properties, each fails to provide detailed clinical information about the patient's symptom severity[24].

Storch and colleagues found problems in convergent validity when developing the Y-BOCS-II[38]. Later, Wu and colleagues performed another study on the psychometric properties of the Y-BOCS-II and found problems in divergent validity with depressive symptoms. However, these two studies used different instruments to assess convergent and divergent validity. Storch and colleagues used the Obsessive-Compulsive Inventory Revised to assess convergent validity, the Penn-State Worry Questionnaire to assess divergent validity with anxiety symptoms and the Inventory of Depressive Symptomatology Self-Report to assess divergent validity with depressive symptoms[38]. The study by Wu and colleagues used the NIMH-GOCS to assess convergent validity, the DASS-anxiety to assess divergent validity with anxiety symptoms, the DASS-depression to assess divergent validity with depressive symptoms and the Barret Impulsiveness Scale to assess divergent validity with impulsivity[41]. The use of a clinician-rated instrument (NIMH-GOCS) to assess OC symptom severity by Wu and colleagues may explain why they found better convergent validity. This problem with convergent and divergent validity is not easy to solve and is an inevitable consequence of the high rates of co-morbidity between these conditions. Nevertheless, I was able to confirm that the Y-BOCS-II is clearly a very reliable instrument, even when analyzing test-retest reliability at an inter-assessment interval that is substantially longer than has been used in previous reports[38,39,41].

In Chapter 2 I asked if the Y-BOCS-II checklist and interview, the gold-standard instrument used to assess the severity of OC symptoms, is a valid measure to diagnose OCD. I established, to our knowledge for the first time, the Y-BOCS-II cut-off score (total score>13) which has the better performance (sensitivity=84.6%; specificity=97%) in distinguishing OCD patients from healthy subjects and from subjects with mood and anxiety disorders. Although the criterion-related validity was very good, some of the problems in construct validity which were previously described in the literature were confirmed in our sample. Specifically, I found significant positive correlations with trait anxiety (Pearson's *r*=0.68), state anxiety (Pearson's *r*=0.43) and depressive symptoms (Pearson's *r*=0.57), suggesting only moderate divergent/discriminant validity.

It is also important to mention the main limitations of our Y-BOCS-II study. First, I did not assess inter-rater reliability. Also, although I established a cut-off with a high sensitivity and high specificity in identifying patients with OCD, a potential criticism that can be done is that using a psychometric instrument to identify a disorder defined by another psychometric instrument (in this case the Structured Clinical Interview for the DSM) is circular. Another major limitation was the small sample size in the clinical control group. In conclusion, I have successfully translated and validated the Y-BOCS-II for the Portuguese adult population. I demonstrated that the Y-BOCS-II is an instrument with high reliability for assessing OC symptoms severity and can be used to identify OCD with high sensitivity and specificity. However, I also confirmed its previously described problems in construct validity.

5.2 Model-free precedes model-based control in uninstructed sequential action choice in healthy subjects

Trouble in linking symptoms with brain function is not an exclusive problem for OCD, it is transversal to all psychiatry [283–287]. Several authors have voiced concerns that psychiatry research has even experienced a stagnation due to the lack of understanding of the neurobiological underpinnings of disorders which are defined phenomenologically [288]. As application of computational methods to psychiatry research has shown great promise in establishing a link between phenomenological and pathophysiological aspects of mental disorders, I used a computational approach in the third chapter of this dissertation.

Computational psychiatry has emerged as a new field in the 21st century[289–292], bridging the gap between psychiatry, mathematical modeling, biophysical modeling and other fields such as neuropsychology or neuroimaging. It seeks normative computational accounts of neural, cognitive and behavioral data and function. These accounts come from the premise that the brain has evolved to solve computational problems. Alan Turing was one of the first to conceive mental function in exactly this fashion: the mind as specific patterns of information processing supported by a particular hardware which is the brain [293]. Computational psychiatry includes two different approaches: data-driven and theory-driven [291]. Data-driven approaches seek to identify disorder-specific features among high-dimensional big data (such as a high number of resting-state fMRI scans). Theory-driven approaches, like the one followed in Chapter 3, develop and test theoretical, often mechanistic, models which try to explain specific phenomena – in this case, learning by interacting with the environment. The field of reinforcement learning is frequently emphasized as an archetype of the success of theory-driven approaches to cognitive science[294]. Computational processes designed by theoretical mathematicians have been imported to model how humans (or other animals) modify their behavior when experiencing rewarding (or aversive) outcomes. By developing and applying a new sequential decision task which, to our knowledge, is the first RL-inspired task with the capacity to differentiate instructed from uninstructed behavior, I have covered an algorithmic account of how the brain controls behavior in distinct circumstances and tested if this control is dysfunctional in patients with OCD.

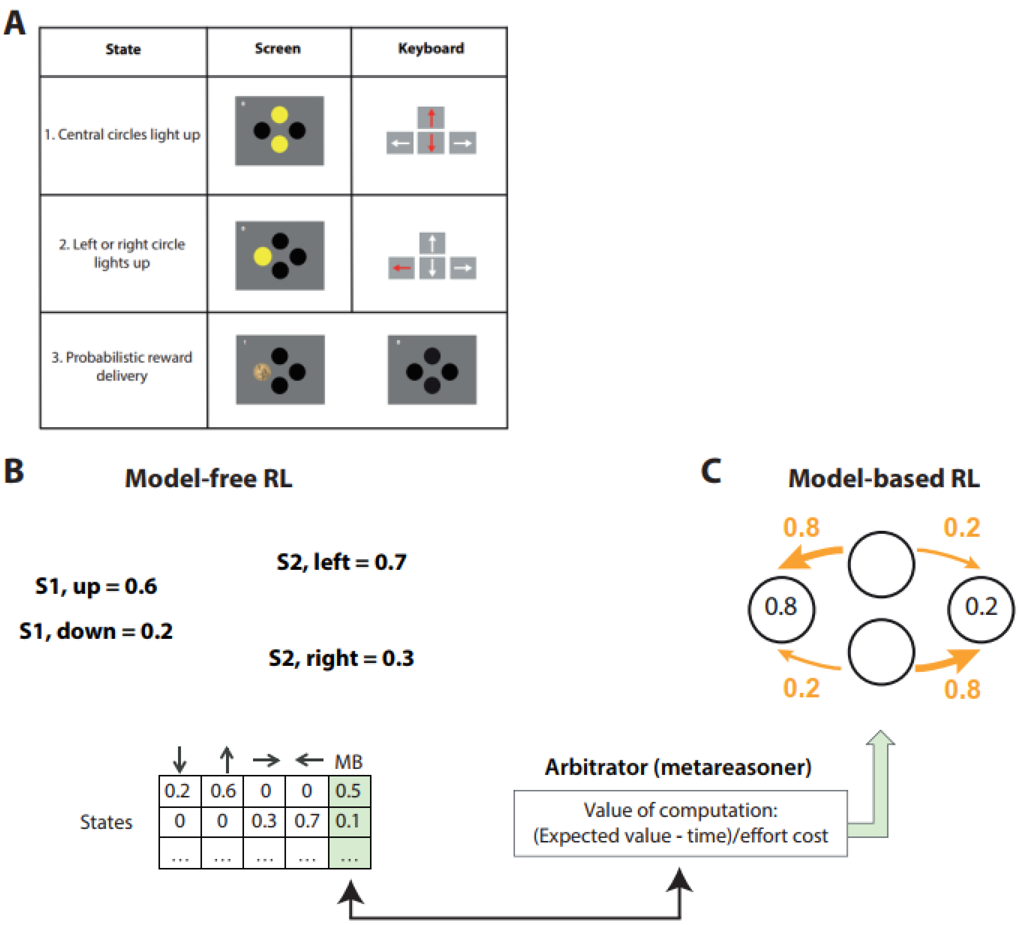
This computational approach allowed for a detailed exploration of uninstructed and instructed learning in healthy subjects. The results of the reduced two-step task in healthy volunteers results that the balance between model-free and model-based strategies in unknown environments is tipped towards higher use of model-free RL, even if the use of model-based RL is advantageous from the beginning. Though model-free RL was the dominant strategy prior to instruction, the influence of model-based RL increased over learning, at least in a subset of individuals. This contrasts with habit formation in classical instrumental learning, where actions are initially goal-directed but become habitual with extended experience [128,152], which most authors interpret as reflecting a transition from model-based to model-free control [158].

Several theories have tried to explain how humans allocate control between these two action control systems [158,295,296]. Following the general RL principle of reward maximization, a potential arbitrator would choose, at each choice point, the controller (model-free or model-based) that is predicted to yield the highest amount of future reward. To attain this objective, at least three different classes of hypotheses have been suggested. In the first of these classes, an arbitrator receives input from both the model-based and the model-free systems in order to decide the balance between their contributions to action [170,183]. The inputs that this arbitrator receives contain information about the quality of the predictions of each system, which can be quantified as a measure of uncertainty. The arbitrator then chooses the system with less uncertainty. Within this class, both systems are engaged at choice point. This class of arbitration rules assumes that there are situations in which the predictions of a model-based system would be worse than the predictions of a model-free system – either due to memory constraints or noise accumulation during computation of model-based action value. This theory captures the key role of uncertainty in the arbitration of goal-directed and habitual mechanisms of choice, and can reproduce the effects of habitization, or the gradual passage from goal-directed to habitual behavior after extensive experience in a stationary environment [152]. This happens because the initial uncertainty of the habitual (model-free) controller,compared to the goal-directed (model-based) one, is higher (as it learns less efficiently from experience) but becomes lower after sufficient learning. This hypothesis assumes that the model-free and model-based controllers are actively engaged in every decision (although ultimately only one of them is selected) and therefore it cannot explain some experimental findings, such as the vanishment of hippocampal forward sweeps (putatively associated with model-based computations) that occurs with habitization in rodent maze tasks [297]. Also, and perhaps the most important drawback of this account, is that it does not consider that model-based computations may have costs, linked to the cognitive effort due to planning [228] and to the temporal discounting of rewards due to the time required for planning [298].

The second class of arbitration rules that was proposed has into account that model-based computations have a high cost (due to its slowness in performing action value calculations or cognitive load) [295]. Based on the assumptions that goal-directed behavior is flexible but slow and habitual behavior is fast but inflexible, and using the computational theory of RL, Keramati and colleagues proposed a normative model for arbitration between the two systems that tries to make an optimal balance between search-time and accuracy in decision making. According to this approach, the model-based system has access to near-perfect information but the value of this information should outweigh its costs to justify using model-based control. Under this arbitration rule, the arbitrator only receives input from the model-free system and the model-based system is only used if the arbitrator decides not to use the model-free system.

Others have suggested the idea of a *mixed instrumental controller* (MIC) which produces patterns of behavior according to a flexible combination of model-based and model-free computations [296]. According to this idea, at decision points, the MIC compares the advantages of model-based computations (in terms of reward) with its costs, performing a sort of cost-benefits analysis. Pezzulo and colleagues propose that this arbitrator calculates the *value of information* of mental simulation (on the basis of uncertainty) and of how much the alternative model-free (cached) values differ against each other, and then compares this value of information against the cost of using model-based RL (in terms of cost and time) [296]. The consequence of this is that model-based control is activated only when necessary.

These three hypothesis for arbitration have recently converged under the unifying account of *computational rationality* [228]. It states that the intelligent brains make computations with algorithms, representations and architectures that are designed to make decisions that lead to the highest utility (immediate worth of states in terms of rewards) while taking into account the costs of computation and views the invocation of the model-based system as a meta-action in which value is estimated by the model-free system (Fig. 29). This is in line with the experimental demonstrations that use of model-based control decreases when cognitive resources are less available [189] and that one of the areas of the cortex which have been associated with the arbitration mechanism (lateral PFC) is also involved in the registration of cognitive demand [183,228].



State	Screen	Keyboard
1. Central circles light up		
2. Left or right circle lights up		
3. Probabilistic reward delivery		

	↓	↑	→	←	MB
States	0.2	0.6	0	0	0.5
	0	0	0.3	0.7	0.1
	…	…	…	…	…

Figure 29. Computational trade-offs in sequential action learning applied to the reduced two-step task.

(A) Reduced two-step task structure. **(B)** A fast but inflexible model-free system stores values for each state-action pair in a look-up table but can also invoke a slower but more flexible model-based system **(C)** which represents the structure of the environment and uses prospective planning (forward search) to construct an optimal sequence of actions. Having a stored value for invoking the model-based system (highlighted in green on the look-up table) is a form of *metareasoning* that weighs the expected value of model-based planning against time and effort costs. Adapted from Gershman et al., 2015 [228]

Although our behavioral experiments do not allow us to identify which type of arbitration is occurring in the human brain, the task structure of the reduced two-step task (higher contrast between common and rare transitions and block-based reward probabilities) favors the use of a model-based system by allowing it to obtain more rewards – a speed/accuracy trade-off arbitrator should then increase the probability of using model-based control when it calculates that a model-free strategy is earning less rewards. The trajectory of arbitration between model-based vs model-free arbitration in the current task likely reflects the dynamic nature of the current task, where ongoing changes in reward probability prevent the model-free system from converging to accurate value estimates and hence dominating behavior late in learning (Fig. 29).

The more complex state space compared with typical instrumental conditioning experiments makes model learning more demanding and hence may increase uncertainty in the model-based system in early learning.

In fact, it has been recently suggested that performance during initial stages of action selection tasks may be primarily based on model-free trial-and-error exploration, with progression towards model-based RL occurring in intermediate stages, as subjects acquire a model of the environment [239]. According to this model, proposed by Bostan & Strick, a later third stage would consist of motor memory learning (Fig. 30). Here the motor memory, or motor learning, in RL terms, means selecting past successful state-action mappings. This three stage model for action learning (1st model-free trial-and-error, 2nd stage: model-based cognitive computations, 3rd stage: motor memory) was proposed after recent fMRI study in healthy humans suggested that distinct brain networks implement different learning strategies to improve performance on an action selection task [299]. Fermin and colleagues in Japan used a “grid-sailing” task that required subjects to move a cursor from an initial point to a goal position in a 5x5 grid. Performance during initial stages of the task was primarily based on trial and error-type exploration (corresponding to model-free RL) and involved a limbic network, including the ventromedial prefrontal cortex (PFC), ventral striatum and posterior cerebellum. As learning progressed and subjects acquired a model of the environment (corresponding to model-based RL), the site of activation shifted to an associative (cognitive) network, including the dorsolateral PFC, dorsomedial striatum and lateral posterior cerebellum. Finally, with extensive experience, performance relied on motor memory, and the site of activation shifted to a motor network, including the supplementary motor area, putamen and anterior cerebellum. Our results of initial reliance on model-free RL and later appearance of model-based control in some subjects are in line with the findings of the Fermin study, although our task was not designed to capture the late stage of motor memory. However, a limitation of the Fermin study is that they did not use computational models to assure that different action selection strategies were indeed in use at the different states.

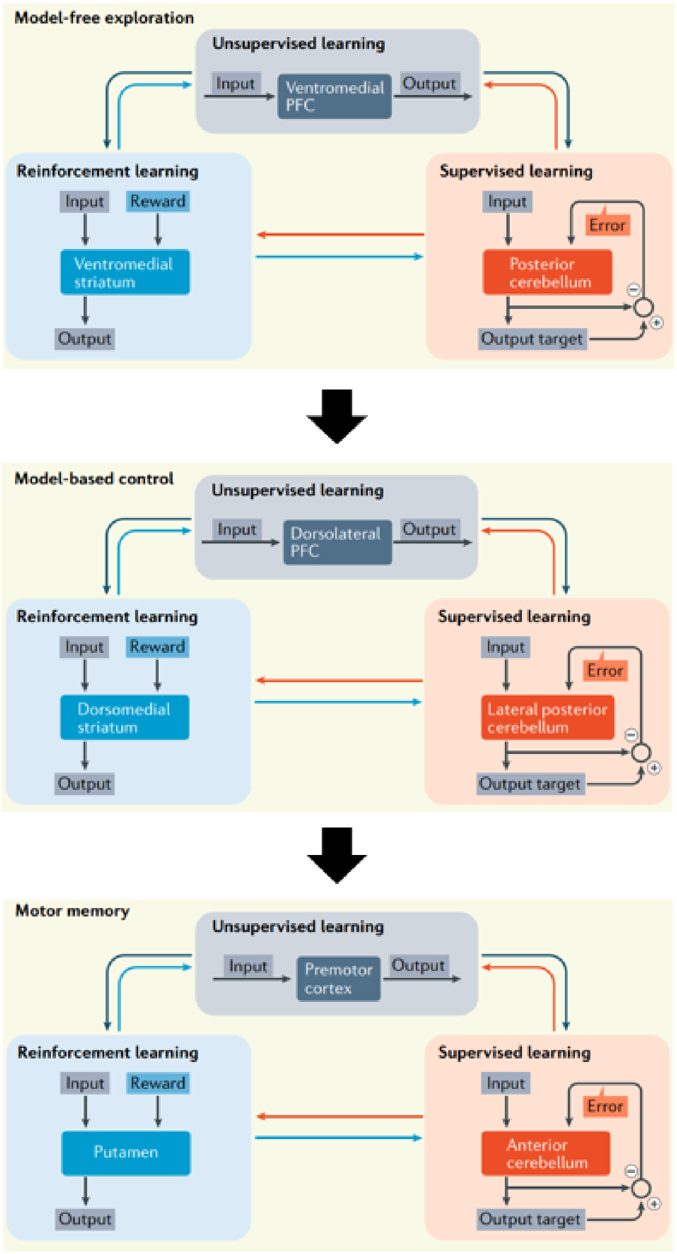


Figure 30. Three distinct action learning strategies with model-free RL preceding model-based RL.

Functionally related brain areas within interconnected networks participate in progressive stages of action learning. Model-free learning through exploration (trial-and-error) involves a limbic network, including the vmPFC, ventral striatum (nucleus accumbens) and posterior cerebellum. Model-based learning occurs later and involves an associative (cognitive) network, including the dorsolateral PFC, dorsomedial striatum (caudate) and lateral posterior cerebellum. Performance based on motor memory involves a motor network, including the supplementary motor area, dorsolateral striatum (putamen) and anterior cerebellum. Imaging data [299] suggests that as learning progresses, the sites of activation shift in a topographically organized fashion from model-free into model-based areas, with motor memory stages appearing later. According to the model depicted in this figure, each stage of the learning progress involves a different set of interconnected basal ganglia, cerebellar and cerebral cortical regions. Adapted from Bostan & Strick, 2018 [239].

Very recently, and relevant to this discussion about a third action learning system, an experiment using a computational analysis of the two-step task asked how humans could learn which components of the environment are important in order to obtain rewards[300]. In the original two-step task, reward probability depends on stimulus identity (images, such as fractals) but not on its spatial localization, making it is possible to ask if behavior shows evidence of assigning value to outcome-irrelevant spatial-motor aspects of the task. This is important because previous studies with the same task only considered outcome-relevant model-free representations and so it was not clear how and if a model-free system could learn which features of the task are relevant for reward prediction and which features are not. The authors found that healthy humans assigned value to spatial-motor representations which were irrelevant for the prediction of reward (e. g. the fractal appearing on the left or on the right side of the screen) and that these representations had effect on behavior[300]. Also, individuals who were more model-based were less prone to this “motor model-free” learning[300]. In our task, the spatial features of the environment (e.g. localization of the circle which lit up or which did not light up) are all relevant for reward prediction, making learning implemented by a typical “stimulus model-free” and potential learning by this newly described “motor model-free system” converge to the same behavior.

In summary, I have shown that in domains where humans lack prior knowledge, model-based RL is slow to develop and behavior relies mosty in model-free control. Next, I explored the impact of explicit knowledge on RL strategies using the same task.

5.3 Diverse effects of explicit knowledge on model-based and model-free reinforcement learning

The reduced two-step task revealed that instructions about task structure in a sequential decision paradigm led to a dramatic increase in the use of model-based RL in healthy subjects (Fig. 31). Several studies have previously explored the impact of explicit instructions in general human behavior using other types of tasks [193,194,301,302]. Curiously, one of the first classical studies on this topic was performed in psychiatric inpatients [301]. The authors found that a reinforcement procedure (e. g. getting candy if they picked up their cutlery after meals) was only effective if it was accompanied by verbal instructions. Their observations also suggested that instructions only had a lasting effect if they were followed by reinforcement. Kaufman and colleagues took a different approach in 1966, closer to modern studies, using students in a laboratory setting [193]. Participants were exposed to a variable-interval (VI) schedule of reinforcement (in which reinforcement was given to the response after an unpredictable amount of time had elapsed) to obtain small monetary rewards and one group was given the correct information that money would be given in a VI basis while two other groups were given incorrect information – one of the groups was told that their actions would be reinforced according to a fixed-interval (FI) schedule (in which the first response was rewarded after a specific and fixed amount of time had elapsed) and the other group was told that their actions would be reinforced according to a variable-ratio (VR) schedule (in which responses are reinforced after an unpredicted amount of responses). Participants in the VR-instructed group responded at high rates, participants in the FI-instructed group responded at low rates and participants in the VI-instructed group responded at intermediate rates. These results led the authors to conclude that instructions – even if inaccurate – exerted powerful influences over rates of response and those influences outweighed the influences of the reinforcement contingencies which were present in the operant conditioning paradigm per se. The same group published another study afterwards, in which participants were trained with five different fixed-interval schedules of reinforcement [194]. Giving subjects information about the contingencies made them respond appropriately to the reinforcement schedules and subjects not provided with information about the schedules responded at very high rates, independently of the schedule. Others have used avoidance schedules to demonstrate that when instructions are incongruent with the reinforcement learning schedule, the type of behavior depends on whether subjects incur in a monetary loss – if subjects do not contact with monetary loss, their behavior follows the instructions; if subjects have contact with monetary loss, their behavior matches the reinforcement schedule [302].

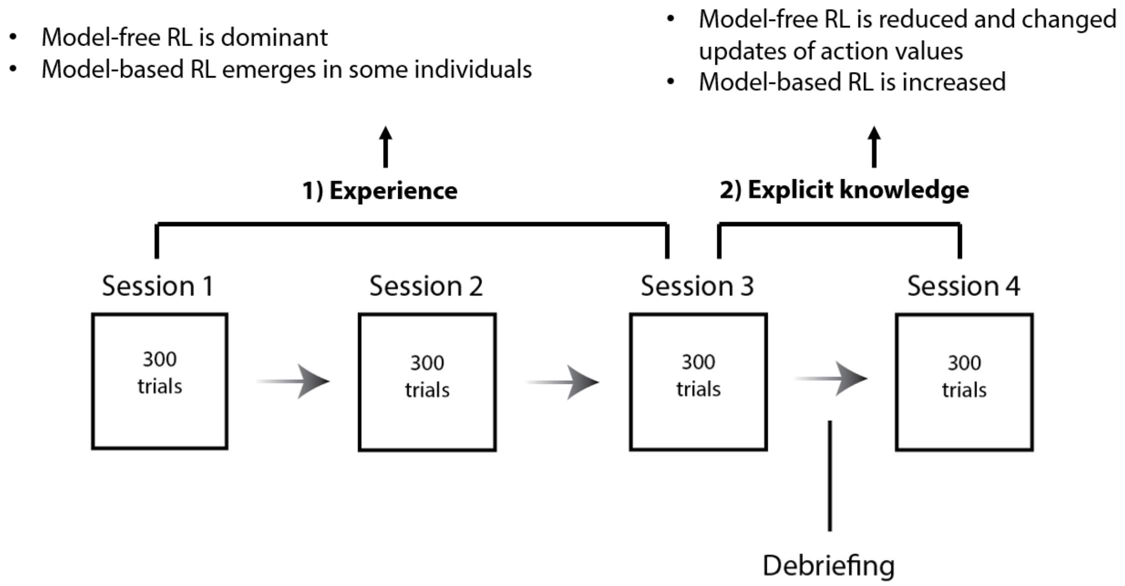

• Model-free RL is dominant
• Model-based RL emerges in some individuals
• Model-free RL is reduced and changed updates of action values
• Model-based RL is increased
1) Experience
2) Explicit knowledge
Session 1
Session 2
Session 3
Session 4
300 trials
300 trials
300 trials
300 trials
Debriefing

Figure 31. Summary of findings in healthy volunteers.

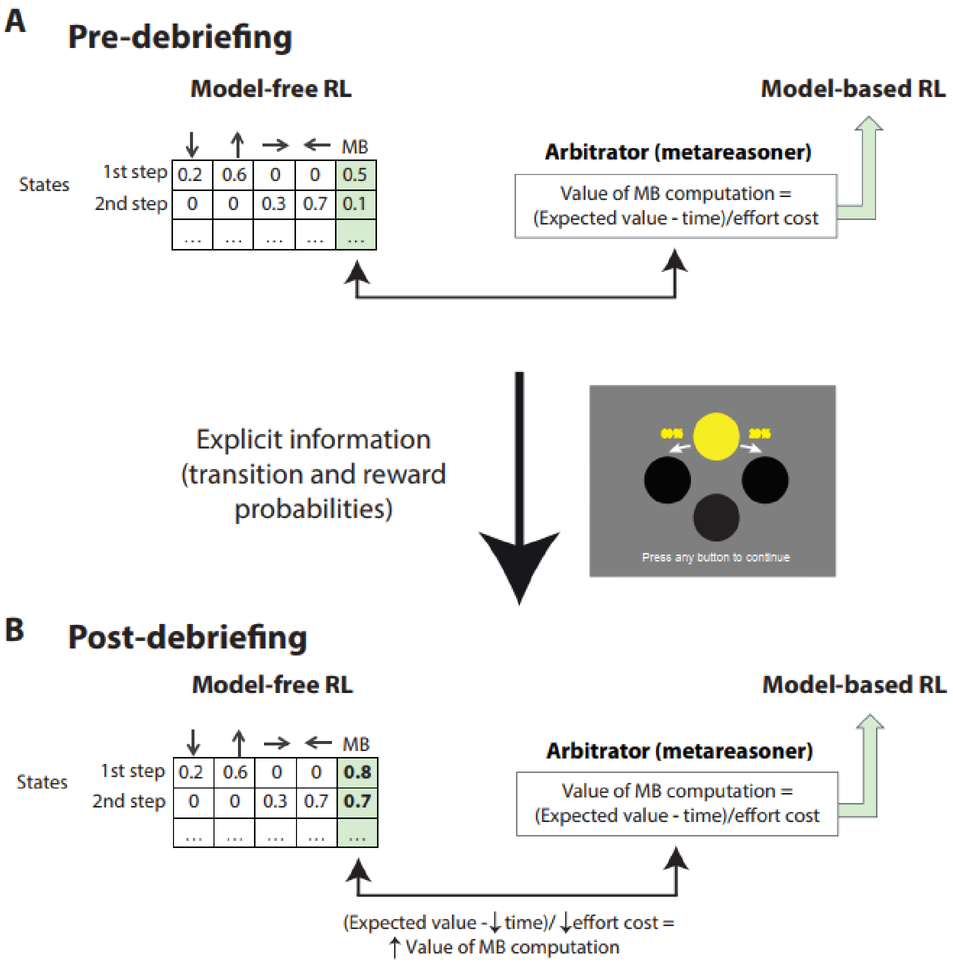
The reduced two-step task allowed, for the first time, to isolate the effects of uninstructed experience and the effects of explicit knowledge in sequential action choice.

More recently, Doll and colleagues have modified a well-known probabilistic decision task to give subjects incorrect information that one of the stimuli is best or worst [198]. Participants' behavior depended on the instructions both initially and after extensive training and the authors tested two possible models (first using neural networks, then using mathematical Q-learning models and then using a Bayesian approach) for the effect of instruction. A central finding from this work is that advice does not just change subjects' initial estimates of how good or bad options are, but also modifies subsequent learning by up-weighting outcomes consistent with advice and down-weighting inconsistent outcomes. Whether such confirmation bias effects extend to task structure learning in addition to simple reward learning is an open question for further work. Neuroimaging has also started to provide mechanistic insights into instruction effects on reward and aversive learning, finding that instruction changes responses to outcomes in striatum and VMPFC/OFC, potentially mediated by instructed knowledge represented in DLPFC [197,200,240]. Our task provides a potential tool for extending such mechanistic investigation of instruction effects into the domain of task structure learning in model-based control.

Our findings are in line with the previous literature regarding the powerful effect that instructions have on human behavior. But I leveraged on previous studies by: 1) using a task that allows to study sequential action selection and 2) implementing a design which allows to analyze computational aspects of uninstructed and instructed behavior in separate. The boost in model-based RL (and the decrease in model-free RL) after the debriefing was in line with our predictions. According to an arbitration mechanism that takes into account the time and effort (i.e., the cost) needed for model-based computations, the value for invoking the model-based system can be calculated based on:

$$Vmb = \frac{Expected\ value - time}{Cost}$$

Receiving information about the transition probabilities and the reward probabilities should decrease the time and the effort cost for model-based computations. That should increase the value for invoking the model-based system and, consequently, its use for control (Fig. 32).


A
Pre-debriefing
Model-free RL
Model-based RL
States
Arbitrator (metareasoner)
Value of MB computation =
(Expected value - time)/effort cost
Explicit information
(transition and reward
probabilities)
Press any button to continue
B
Post-debriefing
Model-free RL
Model-based RL
States
Arbitrator (metareasoner)
Value of MB computation =
(Expected value - time)/effort cost
(Expected value - ↓time)/ ↓effort cost =
↑Value of MB computation

States	↓	↑	→	←	MB
1st step	0.2	0.6	0	0	0.5
2nd step	0	0	0.3	0.7	0.1
	…	…	…	…	…

States	↓	↑	→	←	MB
1st step	0.2	0.6	0	0	**0.8**
2nd step	0	0	0.3	0.7	**0.7**
	…	…	…	…	…

Figure 32. Debriefing effects on the arbitration between model-free and model-based control.

A) During uninstructed learning (pre-debriefing) a model-free system stores values for each state-action pair in a look-up table but can also invoke a slower but more flexible model-based system which represents the structure of the environment and uses prospective planning (forward search) to construct an optimal sequence of actions. **(B)** The explicit information provided in the debriefing about transition and reward probabilities should reduce the time and the effort cost for model-based computation, increasing the cached value for invoking the MB system (highlighted in green).

However, our results also revealed two unexpected findings. The first one was that information about task structure also affected model-free action value updates, increasing the influence on first-step action value updates of the second-step state value relative to the trial outcome, as indexed by the RL model's eligibility trace parameter. This is surprising, because there is no obvious normative reason why information about task structure should change the use of eligibility traces. None the less, the effect of instruction on the eligibility trace parameter was robustly significant, and replicated in both clinical groups as well as the healthy controls.

I suggest that this relationship was not in fact mediated by changes in a model-free eligibility trace, but rather by changes in how subjects represented the tasks state-space. The simplest computational treatments ignore a very important aspect that humans and other animals have to deal with when performing any task – how to represent the different states of the environment? Distinct state representations can make a task easy, hard, near-intractable or impossible to solve. A task can be easier to solve if it has a small number of Markov states that an animal is able to represent and accrue value to, especially if those states allow to build a smooth value function [303]. On the opposite spectrum, a task can become impossible to solve if the animal does not include the information that is fundamental for its performance in the state representation [303]. Therefore, in a sequential decision task, subjects must not only maximize reward but also optimize task/state representations.

Typically, when RL is applied to decision neuroscience, the behavioral task is considered to have a fixed set of discrete states known to the subject. While this is likely a reasonable assumption when subjects are explicitly told the tasks structure, in tasks where subjects must learn task structure from experience, the brain must jointly learn the state-space of the environment and the values of states and actions online from complex and often ambiguous sensory data. The 'model' of the environment learned by the brain therefore comprises not just the state-action-state transition model, used in model-based RL, but also beliefs about the set of states that exist and the current state of the environment, used by both model-based and 'model-free' RL. In this account, explicit information provided to subjects explaining that reward probability depended on the second-step state, made these states more distinct or salient in subjects internal task representation, such that they were better able to accrue value, which then drove model-free learning at the first step. This hypothesis can be directly tested by ongoing work combining the task with neuroimaging. Another way of potentially confirming this hypothesis would be to run a version of the task where a fraction of trials would be left vs. right choices. Forcing subjects to make a left vs. right choice could increase their attention to whether left or right was best – if our hypothesis about the state space and use of model-based RL is correct, this manipulation would increase the use of model-based control on regular trials. This could be implemented by having the normal version of the task in 75% of the trials but on 25% of trials, the trial would start with the left and right circles lighting up, making the subject choose between them to gain access to a reward with a probability which would be drawn from the standard block probabilities for the left and right. This manipulation could increase the salience for the two possible second-step states in regular trials, boosting the contribution of model-based control if our hypothesis is correct.

The other unexpected effect of instruction was to increase subjects' tendency to repeat choices, as indexed by the RL models' perseveration parameter. This likely reflects a strategy of repeatedly sampling a single option to overcome the tasks stochasticity. Such sampling may be increased by instruction because subjects have a discrete set of hypotheses (left is good or right is good) that they are deciding between, potentially increasing the perceived value of repeated sampling.

A final aspect of our behavioral results that needs an interpretation is the absence of correlation between use of model-based RL and working memory (data not shown). Previous work with the original two-step task has shown that use of model-based RL had a significant positive correlation with performance on a visuospatial working memory task (very similar to the Corsi task used in Chapter 3) [191]. Working memory is defined as a process that provides temporary storage and manipulation of information necessary for complex cognitive tasks as language, comprehension, learning, and reasoning [304]. It enables fast and single-trial learning of any kind of information with two limitations: a capacity or resource limit and a time limit [305]. Although RL and working memory systems are supported by different (although partially overlapping) circuits, the way that they interact is far from solved [306]. It has been proposed that WM may be the same as model-based RL and also that successful WM use in simple environments may inhibit model-free RL [306,307]. For the purpose of this dissertation, which is focused in OCD, having a task which is independent of working memory is an advantage, particularly because working memory deficits have been described in OCD, as mentioned in the Introduction [308,309].

5.4 OCD patients have a bias towards increased model-free RL and use model-based RL after receiving explicit information about task structure

The habit account of OCD – which conceptualizes compulsive actions as hyperactive habits – can provide an interesting explanation for the egodystonic nature of obsessive-compulsive symptoms[310]. If actions are more controlled by antecedent stimuli than by the current goals, patients should feel that their behavior doesn’t make sense or is excessive. However, as reviewed in the Introduction, it is difficult to experimentally induce habits in humans [165]. Failures in outcome devaluation paradigms in compulsive individuals may reflect dysfunction in goal-directed control, rather than overactive habit learning [165]. A true tendency for enhanced habits has never been measured in OCD [129,165] – only inferred from deficits in goal-directed behavior in the slips-of-action test of the Fabulous Fruit Game [113] or shock avoidance paradigms [311], which are very liberal and questionable adaptations of instrumental conditioning paradigms used in rodents [129,165]. The two-step task has provided the strongest evidence for an unbalance between action learning systems, with OCD patients favoring model-free over model-based learning [184]. However, it is not known if this is due to a hyperactive model-free system, an underactive model-based system or both. Also, as the two-step task needs explicit instruction, the differences between OCD patients and healthy controls may be related to working memory deficits, which have been described in OCD.

Our task design allowed to separate uninstructed from instructed behavior in a sequential decision task. The computational analysis allowed to isolate the specific strength of model-based control and model-free control, instead of relying on a weighting parameter capturing the balance between them (and an inverse temperature parameter controlling stochasticity) as most studies have done[181,184]. Our results show, to our knowledge for the first time, that OCD patients have a tendency to increase use of model-free RL instead of model-based RL when learning exclusively by interaction with the environment (Fig. 33). Interestingly, I found that OCD patients have higher loading on the model-fitting parameter reflecting model-free RL strength when comparing later training (session 3) with early training (session 1). This early vs. late difference in the model-free parameter was not present in the healthy volunteers or in the group of patients with mood and anxiety disorders. Although the “session x group” interaction forthe model-free parameter in OCD vs. healthy controls did not reach significance, there was a trend in that direction. These results suggest that increased use of model-free RL may have some specificity for OCD.

Our results also show that OCD patients have difficulty in learning a model of the dynamics of the environment exclusively by interacting with it. While healthy subjects show an increase in their loading on the “transition x outcome” predictor from the first to the third session, OCD patients fail to show the same increase. Nevertheless, although there is a significant “session x group” interaction when comparing OCD with healthy controls, the difference in the absolute values of the predictor loadings at session 3 does not reach significance. Moreover, the mood and anxiety group also did not show a difference in use of model-based control between the first and the third session, just like OCD patients, although their variability in use of model-based RL at session 3 was much higher than in the OCD group. Integrating this finding with the evidence found in the literature that deficits in model-based RL are present in other disorders such as methamphetamine addiction, binge eating, alcohol dependence and schizophrenia, it should be concluded that the model-based deficit has a low sensitivity and particularly a low specificity for being used as a consistent marker for OCD [184–186].

It is important to note that shifts from MB to MF across several psychiatric disorders have often been a result of reductions in the MB component, rather than more prominent MF components, both neurally and behaviorally [235,312]. This led a number of authors to raise the possibility that the MB to MF shift can be a result of nonspecific impairments in executive function [190,231] or stress [188] affecting resources for MB computations. In our task, although OCD patients showed problems in increasing their use of model-based RL from direct task experience, their use of model-based control after the debriefing was not different from that of healthy controls. This suggests that OCD patients retain the capacity to use model-based action control in domains where they had prior experience.

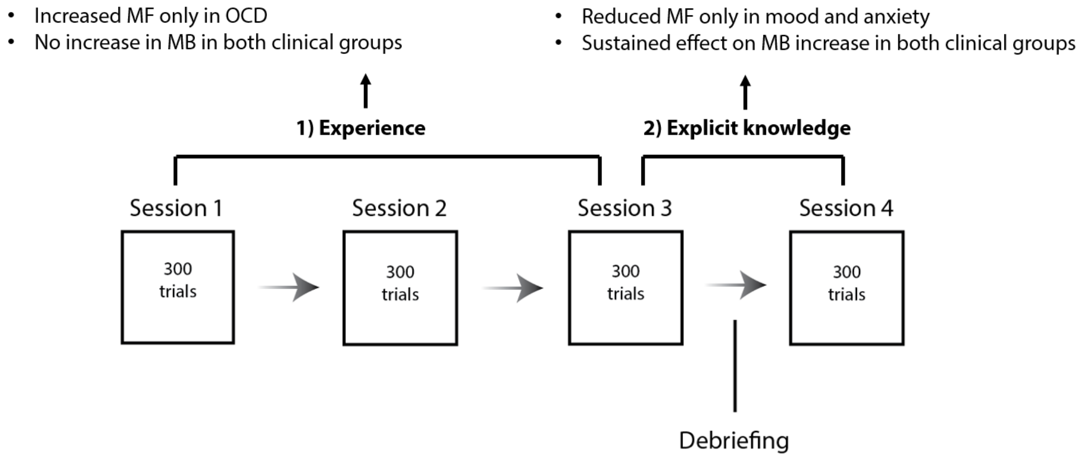


Figure 33. Summary of effects in clinical groups.

Our task design and computational models allowed to isolate the contributions of model-free and model-based RL to uninstructed and to instructed sequential action choice in OCD patients and in patients with mood and anxiety disorders. MB = model-based control. MF = model-free control.

The fact that OCD patients behave like the healthy controls after the instructions in our task may seem contradictory with previous reports describing impaired model-based performance in the original two-step task, where instructions are provided before behavior is analyzed [184]. This may suggest that either OCD patients were unable to understand the instructions of the original two-step task or that they were unable to transfer the information they were given into model-based planning. Also, I propose that previous experience in a task environment may facilitate the use of explicit knowledge by OCD patients. Yet, I did find a deficit in model-based control prior to the debriefing in OCD patients. Still, this deficit was also present in the group of patients with mood and anxiety disorders, demonstrating that it is not specific of OCD, possibly related to common mood and anxiety symptoms, or their underlying mechanisms.

On the other hand, our results suggest that a tendency to increase use of model-free RL during uninstructed learning may be more specific of OCD patients, as this increase was not found in healthy volunteers or in the group of patients with mood and anxiety disorders. Also, the debriefing led to a non-significant difference between pre-debriefing and post-debriefing model-free parameter (although there was a decrease at trend level) in OCD patients, while the other two groups had a very significant decrease in the same parameter. Unfortunately, between-group comparisons also remained at trend level, suggesting that a study with larger groups or with some task modifications may be needed.

Our assessment of action control was transversal but, using the original two-step task [313] a group of our collaborators at the New York State Psychiatric Institute has recently made the first longitudinal assessment of model-based vs. model-free RL in

OCD. They were interested in testing if disruptions in model-based task performance could be state-dependent: an epiphenomenon, present only during, and perhaps resulting from, the presence of acute OCD symptoms [313,314]. They found that OCD symptoms significantly improved following CBT but model-based performance was unaffected by treatment [313]. This led the authors to suggest that deficits in model-based/goal-directed behavior could be a trait or a risk factor for obsessive-compulsive symptoms. However, the limitations imposed by the two-step task and by the restricted computational analysis that was used did not allow to isolate uninstructed action learning nor specific changes in the model-free RL system. In our task, it would be interesting to test if pre- or post-debriefing behavioral measures are associated with better response to CBT. This can in fact be tested in the future since most of the OCD patients recruited in the New York center were treatment-naïve and some of them began CBT after performing our task. Using a behavioral measure (post-debriefing loading in the MB or MF parameters or difference between pre- and post-debriefing) as a predictor of treatment response or as a factor for treatment selection would be a huge step for OCD treatment.

The relationship between obsessions and compulsions – and between them and action control – is also a matter of debate. Children diagnosed with OCD often deny that their compulsions are driven by obsessive thoughts or by anxiety [315,316]. Gillan and colleagues have provided evidence for post-hoc rationalizations of compulsive-like behavior in OCD patients and it has been proposed that the primary phenomenon in OCD could be a tendency to perform compulsive actions, opposing the prevailing conceptual model which posits that obsessions drive compulsive rituals [311,317]. This has even lead to a suggestion for a renaming of OCD into COD (compulsive-obsessive disorder) [318] , but others have opposed this idea using three main arguments: first, that post-hoc rationalizations lack the severity and complexity of the irrational beliefs typically present in OCD; second, that the COD hypothesis does not explain why OCD patients have negative intrusive thoughts instead of neutral or positive intrusive thoughts; third, that intrusive thoughts are also present in other conditions such as social anxiety disorder or post-traumatic stress-disorder that lack compulsive behaviors [314]. This discussion highlights the extraordinary complexity of integrating phenomenology with behavioral data, which we have reviewed elsewhere[319].

Several authors have proposed that the problem in OCD may be specifically related to avoidance habits [311,320]. These authors argue that compulsions in OCD are avoidant, in the sense that they are made to avoid a negative outcome, instead of appetitive, in the sense of being made to obtain a positive outcome. Gillan et al used a shock avoidance paradigm in which participants could avoid a mild electric shock by pressing the correct foot pedal in response to a warning stimulus. After overtraining, the balance between goal-directed and habitual was tested via an instructed outcome devaluation procedure – one of the subjects' wrists was disconnected from the electric stimulator (devalued) while the other remained connected (valued). I question if this type of outcome devaluation isn't in fact a contingency degradation, although ultimately this would be indifferent in terms of behavior. After the instructed devaluation, the number of (unnecessary) responses to the safe (devalued) stimulus was measured. Following overtraining, OCD patients showed greater avoidance of the stimulus which was no longer predictive of a shock, leading the authors to conclude that OCD patients have enhanced avoidance habits [311]. Interestingly, patients that developed avoidance habits showed hyperactivity of the caudate nucleus – a region classically associated with goal-directed behavior [320]. As a future direction, it would be interesting to test behavior of OCD patients in our uninstructed task if the outcome was an aversive stimulus instead of a monetary reward.

A limitation of the habit account of OCD is that it fails to characterize one of its clinical aspects: the high levels of anxiety experienced by OCD patients [321]. A possible bridge between action and anxiety in OCD can be a dysregulation in pavlovian learning. Pavlovian learning is a mechanism by which an animal can learn to make predictions about when biologically significant events are likely to occur, and in particular to learn which stimuli tend to precede them [136]. The contingency that controls Pavlovian learning is the contingency between the stimulus and the outcome instead of stimulus-response association (as in habits) or action-outcome contingency (as in goal-directed behavior) [134]. Behaviors implemented through pavlovian mechanisms are more flexible than reflexes (the simplest type of behavior) in that the moment when behaviors are emitted is shaped by predictive learning, but they are also inflexible since the responses themselves are stereotyped and non-modifiable [134]. Basic emotions such as fear are learned through pavlovian mechanisms. Two studies have found abnormal fear extinction in patients with OCD[322,323]. Another recent study looked at how OCD patients adjusted their behavior to reversals in pavlovian contingencies[324]. The authors found that OCD patients fail to flexibly update their fear responses (measured by skin conductance changes) despite normal initial fear conditioning. This inability to update their threat estimations was correlated with increased BOLD activity in the prefrontal cortex (specifically, in the vmPFC) during the initial stages of learning [324]. These three studies support the idea of aberrant pavlovian information processing in OCD [322–324]. Another interesting future direction of our work would be to modify the reduced two-step task by showing OC-related stimuli instead of rewards.

Reinforcement learning has also been used very recently to try to characterize avoidance in anxiety disorders [325]. These authors operationalize avoidance as a prepotent bias towards withholding actions (inhibition, i.e. “no-go”) when facing potentially negative outcomes [325]. This is a powerful prepotent bias which has been repeatedly observed in animals and in humans and that can have influence in instrumental behavior – a process known as pavlovian-to-instrumental transfer, or PIT [132]. The interaction between anxiety and model-based control in the two-step task has also been directly tested and the results are contradictory. In a 2015 study, the authors reported that experimentally-induced anxiety (via CO2 inhalation) was associated with reduced model-based learning [326]. However, in a preprint already available (not peer-reviewed yet), the same group reports that the same anxiety-inducing manipulation had no effect in model-based RL [327]. In the group of patients with mood and anxiety disorder, I found a relatively poor pre-debriefing model-based control (which did not increase with experience) and normal post-debriefing model-based control. This is in line with a previously published paper in which patients with social anxiety disorder (SAD) had deficits in goal-directed behavior (measured in an instrumental conditioning task) [328]. In this last study, the lack of outcome devaluation was associated with greater symptom severity and poorer response to therapy [328]. A report of increased model-based control in SAD has recently become available (as a preprint, not peer-reviewed yet)[329], but studied online participants only, not clinically diagnosed with SAD[329].

Both our clinical groups (OCD and mood and anxiety disorders) had higher levels of depressive symptoms, has assessed by the BDI-II. Major depressive disorder (MDD) has been proposed to be associated with impaired capacity for reward-based learning [330,331]. Its has also been proposed that learning from negative feedback (punishment) may be impaired in MDD patients who are treated with SSRI's[332,333]. Regarding computational RL, MDD has also been associated with a bias towards negative paths in the mental simulations which are used for model-based planning [334]. In fact, it has been shown in a decision task which was able to highlight specific paths in model-based reasoning that healthy individuals typically disregard branches of a MB decision tree which predict negative outcomes and that this "pruning" was correlated with subclinical depressive symptoms [335]. Our task involved a very small decision tree for this pruning mechanism to be important. Nevertheless, given that OCD has a high-comorbidity with depression – and that, as I have shown, the Y-BOCS-II has only a moderate divergent validity when tested against the BDI – this bias could also be present in OCD patients, but was never tested with the adequate task. This could be related to the phenomenon of the typical content of obsessions being potential negative outcomes, something that has not been adequately explained or formalized.

Our study was not free of limitations. Sequential decision tasks involve trade-offs between the need for determinism, which makes it worth for participants to engage in the task, and stochasticity, which allows behavioral strategies to be discriminated. Our task included blocks with neutral reward probabilities (equal on the left and on the right) to make less obvious to subjects that there are just two possible configurations of the reward probabilities, to make it less likely they will learn some model-free strategies (latent state representations) which may mask as model-based control, as it has been described in simulations with the original two-step task[238]. However, after analyzing the complete behavioral results, in which model-free control dominates (at least during uninstructed experience), it could be advantageous to have a version without neutral blocks in order to facilitate learning. Our preliminary results in the fMRI version (described in Chapter 4 and in the next section), which did not include neutral blocks, suggest that this could be a reasonable strategy to facilitate use of model-based control. Another potential limitation was the absence of application of the changing (transition probabilities) version in the clinical groups. It would be interesting to test the specificity of the model-free strength in a task with a more complex structure.

The problem with comorbidity between OCD and symptoms of another conditions such as anxiety or depression seems to pose a problem when trying to identify an instrument which discriminates between them. Coupling neuroimaging with sequential decision tasks may be an answer to this problem, and I will now discuss the preliminary results from the fMRI experiment.

5.4 Instructions modify brain activity during a sequential decision task

The fourth chapter of this thesis represents the first step of a project aiming at understanding brain dynamics underlying sequential action choice in healthy humans and in OCD patients. The dynamics of the circuits which seem to implement model-based and model-free processes in the healthy brain are not perfectly clear yet, although a limited number of studies suggest that both RL systems are implemented in the same areas, with other studies suggesting distinct but overlapping circuitry (see Fig. 22). Importantly, to our knowledge, no published study has analyzed functional brain activity in OCD patients performing a sequential decision task.

After running three pilot versions of a protocol designed to isolate the three main events of interest in the reduced two-step task (choice; transition; reward) I came to a final version in which participants performed 200 uninstructed trials, before a debriefing was presented inside the scanner, followed by 100 post-debriefing trials. The results extracted from our preliminary data are very promising. Importantly, I demonstrate that, with the version of the task that was adapted for use inside the fMRI scanner, I observe behavioral effects which are very similar to the effects which were present in the (laptop) version of the task and which formed the core of Chapter 3. The most clear and crucial effect is the increase in model-based RL and the decrease in model-free RL when pre-debriefing sessions are compared with post-debriefing sessions. I also show that choice, transition and reward events activate different brain areas. In healthy subjects, in agreement to what would be predicted, the action of pressing for the upper or lower circle activated the motor cortex and the reward delivery activated the ventral striatum, across all sessions. The activation of the motor cortex is not particularly relevant for the context of action learning as it simply reflects action implementation, however, it confirms that the timeseries I am using in the GLM is perfectly time-locked with the stimuli and the behavioral events.

The ventral striatal activation for reward, on the other hand, is extremely relevant for the context of action learning, as this is the fundamental area for reward prediction and one of the main targets for dopaminergic projections from the VTA[274,336]. In fact, it was precisely in the dopaminergic VTA neurons which project to the ventral tegmental area that Schultz and colleagues observed signals which fluctuated in parallel with the reward prediction error, an index of surprise which is part of temporal-difference model-free RL algorithms [168,175]. Interestingly, Daw and colleagues found that BOLD signals in the ventral striatum correlated not only with model-free reward prediction errors but also with a state prediction error, which can only be used by a model-based system[181]. In the same study, the same correlation with both model-free and model-based predictors was also observed in the medial prefrontal cortex. These findings were unexpected at the time and have been hard to reconcile with the dual-system theory and the empirical evidence for separate action control systems that comes from animal studies[132]. An initial explanation for this findings was the suggestion that the model-based system could be training the model-free system by simulating experiences offline[132,181]. Another possible way of interpreting Daw's findings, which was advanced very recently by Matthew Botvinick and colleagues, regards the prefrontal cortex, the striatum and the thalamus as a recurrent neural network[337]. According to these authors, dopaminergic projections from the VTA could be conveying the reward prediction error into both the ventral striatum and the prefrontal cortex, with the projections into the prefrontal areas serving to drive learning by adjusting the synaptic weights in this area[337]. In our task, the transition from the first- into the second-step is a crucial element for learning the structure of the task and an area involved in model-based learning should be active during this transition event. Although I did not find a particular brain region to be more active across all runs/sessions, I found that the debriefing increased BOLD activity during transition events in a cluster extending from the paracingulate cortex into the frontal pole. These prefrontal areas, particularly the frontopolar cortex, have been previously associated with arbitration mechanism between model-based and model-free RL[183]. The main limitation of the current fMRI experience is the sample size (n=9) and the duration of the task inside the scanner (~90 minutes), which may be problematic for OCD patients. I plan to expand data collection until 40 healthy subjects have been included, and I plan to integrate RL-driven predictors into the GLM. Other future directions include using multivariate approaches, such as representation-similarity analysis, which have recently began to be used in fMRI studies to try to understand how different task aspects are represented in the brain[279,280,338].

5.5. Conclusions

In conclusion, I tested, for the first time, criterion validity of the gold-standard instrument to assess the severity of OC symptoms and I found that the best cut-off for the diagnosis of OCD is a total score above 13 points. I also confirmed the problems in construct validity that the Y-BOCS-II presents. Next, inspired by the literature which suggests that deficits in model-based action control may be a marker of OCD, I developed a new sequential decision task – which tackles some of the limitations of the most popular paradigm and that allows to separate the effects of experience in the task from the effects of explicit knowledge. I applied the new task in healthy volunteers, as well as in OCD patients and controls with mood and anxiety disorders. Both clinical groups had trouble in increasing their use of model-based control with experience but were able to use it after being instructed. However, the OCD group was the only group in which uninstructed experience increased their use of model-free RL and in which explicit information did not decrease use of model-free RL, suggesting that an hyperactive model-free system may be more specific of OCD than underactive model-based control. I also found new and unexpected effects of debriefing on model-free action value updates in healthy and clinical groups. Finally, I designed a protocol which shows capacity to measure brain activity in a sequential decision task with uninstructed and with instructed sessions. We found that debriefing modified brain activity in an area extending from the paracingulate cortex into the frontal pole, suggesting that the changes in behavior after explicit information is provided may be driven by a modification in brain activity in medial prefrontal areas. Future directions include combining computational models with behavioral and neuroimaging data, in healthy and clinical populations, in order to attain the goal of finding the first consistent biological marker of OCD.

Supplementary information

Information before task

“You will now play a game in order to gain of as many rewards as possible.

Rewards will be represented in the screen as coins. Every time you get a coin, it will show up in the screen and it will be added to your total number of rewards. The number of coins you get will determine the value of the gift-card that you will receive at the end of your participation.

You will perform 1200 trials and in each trial you can get either one coin or no coin. At the end of those 1200 trials, 400 will be randomly chosen to count the final number of coins.

The minimum amount of money in your gift card will be 10 euros. For each coin that you get above 150 coins, you will get an increase of 20 cents in your gift card. Therefore, if you get 175 coins the amount will be 15 euros, 200 coins correspond to 20 euros and 225 coins correspond to the maximum amount that the gift-card can have, which is 25 euros. Amounts will be distributed rounded to the closer multiple of 5 euros.

At the top left corner of the screen, there will be a coin counter which shows how many coins you got in each session. That number may not have direct correspondence with the final amount, since that amount will be calculated using a random sample of trials.

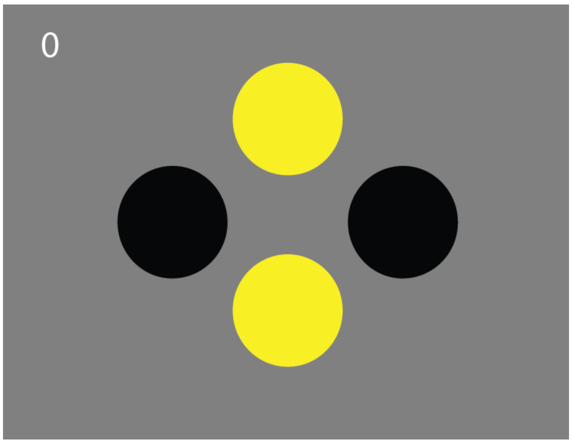
You will play the game using the arrow keys after stimuli show up in the screen.

Each session of the game will last for approximately 15 minutes. Once the session is completed, a sentence thanking you for your participation will show up in the screen. When that screen shows up you should leave the room.

Debriefing – Fixed transition probabilities version

We will now explain the structure of the game.

First the two central circles (upper and lower) are yellow, indicating that you can choose one of them.



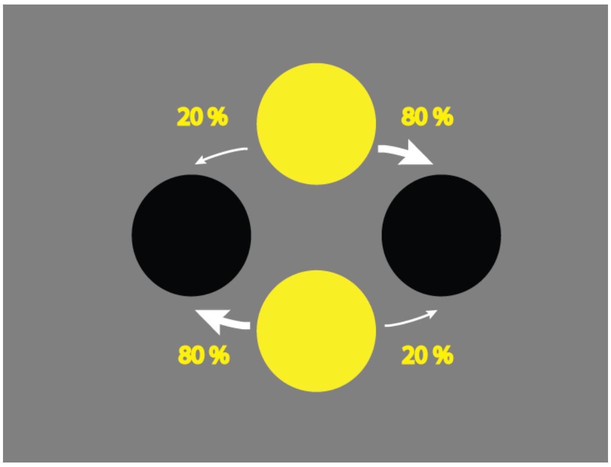
If you press the upper arrow key, you will choose the upper circle. If you press the lower arrow key, you will choose the lower circle.

After you choose the upper or the lower circle, one of the two side circles will light up, i. e., will turn yellow (left or right). After you press the arrow key that corresponds to the lateral circle that lit up (left or right), a coin may or may not appear.

The probability according to which the central circles give access to either one of the lateral circles also follows some rules.

If you choose the upper circle, one of two different things can happen. Most of the times (actually 80% of the times), the right side circle will light up. Rarely, the left side circle will light up.

If you choose the lower circle, most of the times (actually 80% of the times) the left side circle will light up. On the remaining occasions, the right side circle will light up.



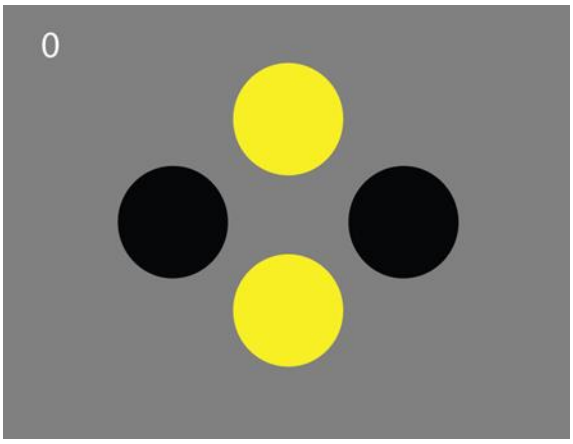
The left and right circles give access to the rewards, which are symbolized as coins. However, the probability of winning a coin is not the equal on the left or on the right: it is always higher on one of the sides. Sometimes it is higher on the left and sometimes it is higher on the right. The side in which that probability is higher changes after 20 or more trials.

You will now play a last session, with the same rules. Good luck!

Debriefing – Changing transition probabilities version

We will now explain the structure of the game.

First the two central circles (upper and lower) are yellow, indicating that you can choose one of them.



If you press the upper arrow key, you will choose the upper circle. If you press the lower arrow key, you will choose the lower circle.

After you choose the upper or the lower circle, one of the two side circles will light up, i. e., will turn yellow (left or right). After you press the arrow key that corresponds to the lateral circle that lit up (left or right), a coin may or may not appear.

The probability according to which the central circles give access to either one of the lateral circles also follows some rules. The game is divided in two types of blocks.

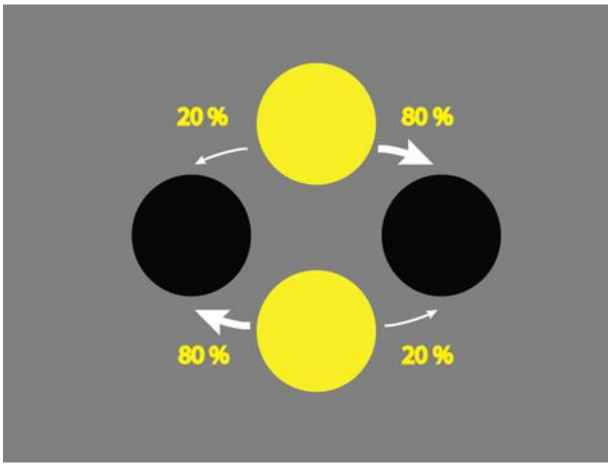
In “A” blocks, choosing the upper circle leads more frequently (80% of the times) to the lighting up of the right side circle. On the other hand, in these blocks, choosing the lower circle, leads more frequently (80% of the times) to the lighting up of the left side circle.

In “B” blocks, choosing the upper circle leads more frequently (80% of the times) to the lighting up of the left side circle. On the other hand, in these blocks, choosing the lower circle, leads more frequently (80% of the times) to the lighting up of the right side circle.

Therefore, in “A” blocks, if you choose the upper circle, one of two things can happen. Most of the times (actually 80% of the times), the right side circle will light up. Rarely (20% of the time), the left side circle will light up.

In these same “A” blocks, if you choose the lower circle, one of two things can happen. Most of the times (actually 80% of the times), the left side circle will light up. Rarely (20% of the time), the right side circle will light up.

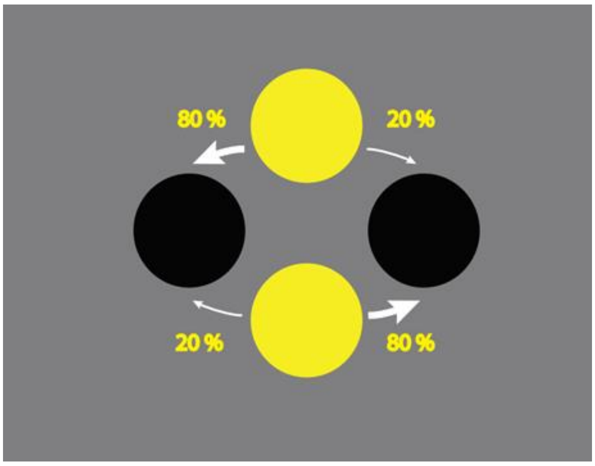
Schematic representation of the structure of “A” blocks:



In “B” blocks, if you choose the upper circle, one of two things can happen. Most of the times (actually 80% of the times), the left side circle will light up. Rarely (20% of the time), the right side circle will light up.

In these same “B” blocks, if you choose the lower circle, one of two things can happen. Most of the times (actually 80% of the times), the right side circle will light up. Rarely (20% of the time), the left side circle will light up.

Schematic representation of the structure of “B” blocks:



“A” blocks and “B” blocks alternate between them after 20 or more trials.

The left and right circles give access to the rewards, which are symbolized as coins. However, the probability of winning a coin is not the equal on the left or on the right: it is always higher on one of the sides. Sometimes it is higher on the left and sometimes it is higher on the right. The side in which that probability is higher changes after 20 or more trials.

You will now play a last session, with the same rules. Good luck!

References

1. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders*. *World Health Organization* (1993). doi:10.4103/0019

2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5). *Diagnostic Stat. Man. Ment. Disord. 4th Ed. TR.* 280 (2013). doi:10.1176/appi.books.9780890425596.744053

3. Rasmussen, S. A. & Eisen, J. L. The epidemiology and differential diagnosis of obsessive compulsive disorder. 144th Annual Meeting of the American Psychiatric Association: Obsessive compulsive disorder: Integrating theory and practice (1991, New Orleans, Louisiana). *J. Clin. Psychiatry* (1992).

4. Kessler, R. C. *et al.* Lifetime Prevalence and Age-of-Onset Distributions of. *Arch Gen Psychiatry* **62**, 593–602 (2005).

5. Caldas de Almeida, J. *et al.* Estudo Epidemiológico Nacional de Saúde Mental - 1.º Relatório. *Lisboa. Fac. Ciências Médicas, da* 60 (2013).

6. Kessler, R. C. *et al.* Lifetime Prevalence and Age-of-Onset Distributions of. *Arch Gen Psychiatry* **62**, 593–602 (2005).

7. Geller, D. A. Obsessive-Compulsive and Spectrum Disorders in Children and Adolescents. *Psychiatric Clinics of North America* (2006). doi:10.1016/j.psc.2006.02.012

8. Ruscio, A. M., Stein, D. J., Chiu, W. T. & Kessler, R. C. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Mol. Psychiatry* **15**, 53–63 (2010).

9. Do Rosario-Campos, M. C. *et al.* Adults with early-onset obsessive-compulsive disorder. *Am. J. Psychiatry* (2001). doi:10.1176/appi.ajp.158.11.1899

10. Mathis, M. A. de *et al.* Gender differences in obsessive-compulsive disorder: a literature review. *Rev. Bras. Psiquiatr.* (2011). doi:10.1590/s1516-44462011000400014

11. Lopez, A. D. & Murray, C. C. J. L. The global burden of disease, 1990-2020. *Nature Medicine* (1998). doi:10.1038/3218

12. Subramaniam, M., Soh, P., Vaingankar, J. A., Picco, L. & Chong, S. A. Quality of life in obsessive-compulsive disorder: Impact of the disorder and of treatment. *CNS Drugs* (2013). doi:10.1007/s40263-013-0056-z

13. Macy, A. *et al.* Quality of life in obsessive compulsive disorder. *CNS Spectr.* (2013). doi:10.1007/978-1-60327-343-5_30

14. Hollander, E. *et al.* Refractory obsessive-compulsive disorder: State-of-the-art treatment. in *Journal of Clinical Psychiatry* **63**, 20–29 (2002).

15. Kaplan, A. & Hollander, E. A Review of Pharmacologic Treatments for Obsessive-Compulsive Disorder. *Psychiatr. Serv.* **54**, 1111–1118 (2003).

16. Pallanti, S. & Quercioli, L. Treatment-refractory obsessive-compulsive disorder:

Methodological issues, operational definitions and therapeutic lines. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **30**, 400–412 (2006).

17. Fineberg, N. A. *et al.* Obsessive-compulsive disorder (OCD): Practical strategies for pharmacological and somatic treatment in adults. *Psychiatry Res.* **227**, 114–125 (2015).

18. Cottraux, J. *et al.* A randomized controlled trial of cognitive therapy versus intensive behavior therapy in obsessive compulsive disorder. *Psychother. Psychosom.* (2001). doi:10.1159/000056269

19. Fals-Stewart, W., Marks, A. P. & Schafer, J. A comparison of behavioral group therapy and individual behavior therapy in treating obsessive-compulsive disorder. *J. Nerv. Ment. Dis.* (1993). doi:10.1097/00005053-199303000-00007

20. Foa, E. B. *et al.* Randomized, placebo-controlled trial of exposure and ritual prevention, clomipramine, and their combination in the treatment of obsessive-compulsive disorder. *Am. J. Psychiatry* (2005). doi:10.1176/appi.ajp.162.1.151

21. Lindsay, M., Crino, R. & Andrews, G. Controlled trial of exposure and response prevention in obsessive- compulsive disorder. *Br. J. Psychiatry* (1997).

22. Nurmi, E. L. & Eyal, R. Pharmacotherapy in the treatment of obsessive-compulsive disorder. in *Obsessive compulsive disorder: etiology, phenomenology and treatment* (ed. Lack, C. W.) (Onus books, 2015).

23. Schuurmans, J. *et al.* The Netherlands Obsessive Compulsive Disorder Association (NOCDA) study: Design and rationale of a longitudinal naturalistic study of the course of OCD and clinical characteristics of the sample at baseline. *Int. J. Methods Psychiatr. Res.* (2012). doi:10.1002/mpr.1372

24. Storch, E. A., Benito, K. & Goodman, W. Assessment scales for obsessive-compulsive disorder. *Neuropsychiatry* (2011). doi:10.2217/npy.11.22

25. Grabill, K. *et al.* Assessment of obsessive-compulsive disorder: A review. *J. Anxiety Disord.* (2008). doi:10.1016/j.janxdis.2007.01.012

26. Goodman, W. K. *et al.* The Yale-Brown Obsessive Compulsive Scale: I. Development, Use, and Reliability. *Arch. Gen. Psychiatry* **46**, 1006–1011 (1989).

27. Baer, L. Factor analysis of symptom subtypes of obsessive compulsive disorder and their relation to personality and tic disorders. in *Journal of Clinical Psychiatry* (1994).

28. Leckman, J. F. *et al.* Symptoms of obsessive-compulsive disorder. *Am. J. Psychiatry* (1997). doi:10.1176/ajp.154.7.911

29. Summerfeldt, L. J., Richter, M. A., Antony, M. M. & Swinson, R. P. Symptom structure in obsessive-compulsive disorder: A confirmatory factor-analytic study. *Behav. Res. Ther.* (1999). doi:10.1016/S0005-7967(98)00134-X

30. Pinto, A. *et al.* Taboo thoughts and doubt/checking: A refinement of the factor structure for obsessive-compulsive disorder symptoms. *Psychiatry Res.* (2007). doi:10.1016/j.psychres.2006.09.005

31. Bloch, M. H., Landeros-Weisenberger, A., Rosario, M. C., Pittenger, C. & Leckman, J. F. Meta-analysis of the symptom structure of obsessive-compulsive disorder. *Am. J. Psychiatry* (2008). doi:10.1176/appi.ajp.2008.08020320

32. Amir, N., Foa, E. B. & Coles, M. E. Factor structure of the Yale-Brown Obsessive Compulsive Scale. *Psychol. Assess.* **9**, 312–316 (1997).

33. McKay, D., Neziroglu, F., Stevens, K. & Yaryura-Tobias, J. A. The yale-brown obsessive-compulsive scale: Confirmatory factor analytic findings. *J. Psychopathol. Behav. Assess.* **20**, 265–267 (1998).

34. Arrindell, W. A., De Vlaming, I. H., Eisenhardt, B. M., Van Berkum, D. E. & Kwee, M. G. T. Cross-cultural validity of the Yale-Brown Obsessive Compulsive Scale. *J. Behav. Ther. Exp. Psychiatry* (2002). doi:10.1016/S0005-7916(02)00047-2

35. McKay, D., Danyko, S., Neziroglu, F. & Yaryura-Tobias, J. A. Factor structure of the Yale-Brown Obsessive-Compulsive scale: A two dimensional measure. *Behav. Res. Ther.* **33**, 865–869 (1995).

36. Deacon, B. J. & Abramowitz, J. S. The Yale-Brown Obsessive Compulsive Scale: Factor analysis, construct validity, and suggestions for refinement. *J. Anxiety Disord.* (2005). doi:10.1016/j.janxdis.2004.04.009

37. Storch, E. A. *et al.* Factor analytic study of the children's Yale-Brown obsessive-compulsive scale. *J. Clin. Child Adolesc. Psychol.* (2005). doi:10.1207/s15374424jccp3402_10

38. Storch, E. A. *et al.* Development and psychometric evaluation of the yale-brown obsessive-compulsive scale-second edition. *Psychol. Assess.* **22**, 223–232 (2010).

39. Melli, G. *et al.* Validation of the Italian version of the Yale-Brown Obsessive Compulsive Scale-Second Edition (Y-BOCS-II) in a clinical sample. *Compr. Psychiatry* **60**, 86–92 (2015).

40. Hiranyatheb, T., Saipanish, R. & Lotrakul, M. Reliability and validity of the Thai Version Of The Yale-Brown Obsessive Compulsive Scale - second edition in clinical samples. *Neuropsychiatr. Dis. Treat.* **10**, 471–477 (2014).

41. Wu, M. S., McGuire, J. F., Horng, B. & Storch, E. A. Further psychometric properties of the Yale-Brown Obsessive Compulsive Scale - Second Edition. *Compr. Psychiatry* **66**, 96–103 (2016).

42. Dollard, J. & Miller, N. E. Personality and Psychotherapy: An Analysis in Terms of Learning, Thinking, Culture. *Am. Sociol. Rev.* (1950). doi:10.2307/2087628

43. Mowrer, O. H. A stimulus-response analysis of anxiety and its role as a reinforcing agent. *Psychol. Rev.* (1939). doi:10.1037/h0054288

44. Salkovskis, P. M. Obsessional-compulsive problems: A cognitive-behavioural analysis. *Behav. Res. Ther.* (1985). doi:10.1016/0005-7967(85)90105-6

45. Bandura, A. & Walters, R. *Social learning theory*. (Prentice Hall, 1977).

46. Lack, C. W. Obsessive-compulsive disorder: Evidence-based treatments and future directions for research. *World J. Psychiatry* (2013).

doi:10.5498/wjp.v2.i6.86

47. Oliveira-Maia, A. J. & Castro-Rodrigues, P. Brain-derived neurotrophic factor: A biomarker for obsessive-compulsive disorder? *Front. Neurosci.* (2015). doi:10.3389/fnins.2015.00134

48. Do Rosario-Campos, M. C. *et al.* A family study of early-onset obsessive-compulsive disorder. *Am. J. Med. Genet. - Neuropsychiatr. Genet.* (2005). doi:10.1002/ajmg.b.30149

49. Grados, M. A., Walkup, J. & Walford, S. Genetics of obsessive-compulsive disorders: New findings and challenges. in *Brain and Development* (2003). doi:10.1016/S0387-7604(03)90010-6

50. Mataix-Cols, D. *et al.* Population-based, multigenerational family clustering study of obsessive-compulsive disorder. *JAMA Psychiatry* (2013). doi:10.1001/jamapsychiatry.2013.3

51. Van Grootheest, D. S., Cath, D. C., Beekman, A. T. & Boomsma, D. I. Twin studies on obsessive-compulsive disorder: A review. *Twin Research and Human Genetics* (2005). doi:10.1375/183242705774310060

52. Hanna, G. L., Fischer, D. J., Chadha, K. R., Himle, J. A. & Van Etten, M. Familial and sporadic subtypes of early-onset Obsessive-Compulsive disorder. *Biol. Psychiatry* (2005). doi:10.1016/j.biopsych.2004.12.022

53. Arnold, P. D. *et al.* Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol. Psychiatry* (2018). doi:10.1038/mp.2017.154

54. Hollander, E., Braun, A. & Simeon, D. Should OCD leave the anxiety disorders in DSM-V? The case for obsessive compulsive-related disorders. in *Depression and Anxiety* (2008). doi:10.1002/da.20500

55. Pauls, D. L., Abramovitch, A., Rauch, S. L. & Geller, D. A. Obsessive-compulsive disorder: an integrative genetic and neurobiological perspective. *Nat. Rev. Neurosci.* **15**, 410–424 (2014).

56. Zai, G. *et al.* Myelin oligodendrocyte glycoprotein (MOG) gene is associated with obsessive-compulsive disorder. *Am. J. Med. Genet.* (2004). doi:10.1002/ajmg.b.30077

57. Jiang, C. *et al.* Association between TNF -α-238G/A gene polymorphism and OCD susceptibility. *Med. (United States)* (2018). doi:10.1097/MD.0000000000009769

58. Stewart, S. E. *et al.* Genome-wide association study of obsessive-compulsive disorder. *Mol. Psychiatry* (2013). doi:10.1038/mp.2012.85

59. Mattheisen, M. *et al.* Genome-wide association study in obsessive-compulsive disorder: Results from the OCGAS. *Mol. Psychiatry* (2015). doi:10.1038/mp.2014.43

60. Chamberlain, S. R., Blackwell, A. D., Fineberg, N. A., Robbins, T. W. & Sahakian, B. J. The neuropsychology of obsessive compulsive disorder: The importance of failures in cognitive and behavioural inhibition as candidate

endophenotypic markers. *Neuroscience and Biobehavioral Reviews* (2005). doi:10.1016/j.neubiorev.2004.11.006

61. Menzies, L. *et al.* Integrating evidence from neuroimaging and neuropsychological studies of obsessive-compulsive disorder: The orbitofronto-striatal model revisited. *Neuroscience and Biobehavioral Reviews* **32**, 525–549 (2008).

62. Kim, J. *et al.* Grey matter abnormalities in obsessive-compulsive disorder: Statistical parametric mapping of segmented magnetic resonance images. *Br. J. Psychiatry* (2001). doi:10.1192/bjp.179.4.330

63. Pujol, J. *et al.* Mapping structural brain alterations in obsessive-compulsive disorder. *Arch. Gen. Psychiatry* (2004). doi:10.1001/archpsyc.61.7.720

64. Valente, A. A. *et al.* Regional gray matter abnormalities in obsessive-compulsive disorder: A voxel-based morphometry study. *Biol. Psychiatry* (2005). doi:10.1016/j.biopsych.2005.04.021

65. Boedhoe, P. S. W. *et al.* Distinct subcortical volume alterations in pediatric and adult OCD: A worldwide meta- and mega-analysis. *Am. J. Psychiatry* (2017). doi:10.1176/appi.ajp.2016.16020201

66. Boedhoe, P. S. W. *et al.* Cortical abnormalities associated with pediatric and adult obsessive-compulsive disorder: Findings from the enigma obsessive-compulsive disorder working group. *Am. J. Psychiatry* (2018). doi:10.1176/appi.ajp.2017.17050485

67. Fitzgerald, K. D. *et al.* Developmental alterations of frontal-striatal-thalamic connectivity in obsessive-compulsive disorder. *J. Am. Acad. Child Adolesc. Psychiatry* (2011). doi:10.1016/j.jaac.2011.06.011

68. Pauls, D. L., Abramovitch, A., Rauch, S. L. & Geller, D. A. Obsessive-compulsive disorder: an integrative genetic and neurobiological perspective. *Nat. Rev. Neurosci.* **15**, 410–424 (2014).

69. Baxter, L. R. *et al.* Cerebral glucose metabolic rates in nondepressed patients with obsessive-compulsive disorder. *Am. J. Psychiatry* **145**, 1560–1563 (1988).

70. Swedo, S. E. *et al.* Cerebral Glucose Metabolism in Childhood-Onset Obsessive-Compulsive Disorder. *Arch. Gen. Psychiatry* (1989). doi:10.1001/archpsyc.1989.01810060038007

71. Abramovitch, A., Mittelman, A., Henin, A. & Geller, D. Neuroimaging and neuropsychological findings in pediatric obsessive-compulsive disorder: A review and developmental considerations. *Neuropsychiatry* (2012). doi:10.2217/npy.12.40

72. Breiter, H. C. *et al.* Functional magnetic resonance imaging of symptom provocation in obsessive-compulsive disorder. *Arch. Gen. Psychiatry* (1996). doi:10.1001/archpsyc.1996.01830070041008

73. Koch, K. *et al.* Aberrant anterior cingulate activation in obsessive-compulsive disorder is related to task complexity. *Neuropsychologia* (2012). doi:10.1016/j.neuropsychologia.2012.02.002

74. Saxena, S. & Rauch, S. L. Functional neuroimaging and the neuroanatomy of obsessive-compulsive disorder. *Psychiatric Clinics of North America* **23**, 563–586 (2000).

75. Milad, M. R. & Rauch, S. L. Obsessive-compulsive disorder: Beyond segregated cortico-striatal pathways. *Trends in Cognitive Sciences* **16**, 43–51 (2012).

76. Kandel, E., Schwarz, J., Jessel, T., Siegelbaum, S. & Hudspeth, A. J. *Principles of neural science*. *MCGRAW-HILL COMPANIES* (McGraw-Hill, 2000).

77. Alexander, G. Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex. *Annu. Rev. Neurosci.* (1986). doi:10.1146/annurev.neuro.9.1.357

78. Perani, D. *et al.* [18F]FDG PET study in obsessive-compulsive disorder. A clinical/metabolic correlation study after treatment. *Br. J. Psychiatry* (1995).

79. Baxter, L. R. *et al.* Caudate Glucose Metabolic Rate Changes with Both Drug and Behavior Therapy for Obsessive-Compulsive Disorder. *Arch. Gen. Psychiatry* (1992). doi:10.1001/archpsyc.1992.01820090009002

80. Saxena, S. *et al.* Localized orbitofrontal and subcortical metabolic changes and predictors of response to paroxetine treatment in obsessive-compulsive disorder. *Neuropsychopharmacology* (1999). doi:10.1016/S0893-133X(99)00082-2

81. Saxena, S. *et al.* Differential cerebral metabolic changes with paroxetine treatment of obsessive-compulsive disorder vs major depression. *Arch. Gen. Psychiatry* (2002). doi:10.1001/archpsyc.59.3.250

82. Freyer, T. *et al.* Frontostriatal activation in patients with obsessive-compulsive disorder before and after cognitive behavioral therapy. *Psychol. Med.* (2011). doi:10.1017/S0033291710000309

83. Saxena, S. *et al.* Rapid effects of brief intensive cognitive-behavioral therapy on brain glucose metabolism in obsessive-compulsive disorder. *Mol. Psychiatry* (2009). doi:10.1038/sj.mp.4002134

84. De Koning, P. P., Figee, M., Van Den Munckhof, P., Schuurman, P. R. & Denys, D. Current status of deep brain stimulation for obsessive-compulsive disorder: A clinical review of different targets. *Curr. Psychiatry Rep.* (2011). doi:10.1007/s11920-011-0200-8

85. Karas, P. J. *et al.* Deep brain stimulation for obsessive compulsive disorder: Evolution of surgical stimulation target parallels changing model of dysfunctional brain circuits. *Frontiers in Neuroscience* (2019). doi:10.3389/fnins.2018.00998

86. Denys, D. *et al.* Deep Brain Stimulation of the Nucleus Accumbens for Treatment-Refractory Obsessive-Compulsive Disorder. *Arch. Gen. Psychiatry* **67**, 1061 (2010).

87. Huff, W. *et al.* Unilateral deep brain stimulation of the nucleus accumbens in patients with treatment-resistant obsessive-compulsive disorder: Outcomes after one year. *Clin. Neurol. Neurosurg.* **112**, 137–143 (2010).

88. Gabriëls, L., Cosyns, P., Nuttin, B., Demeulemeester, H. & Gybels, J. Deep brain

stimulation for treatment-refractory obsessive-compulsive disorder: Psychopathological and neuropsychological outcome in three cases. *Acta Psychiatr. Scand.* **107**, 275–282 (2003).

89. Abelson, J. L. *et al.* Deep brain stimulation for refractory obsessive-compulsive disorder. *Biol. Psychiatry* **57**, 510–516 (2005).

90. Mallet, L. *et al.* Subthalamic nucleus stimulation in severe obsessive-compulsive disorder. *N. Engl. J. Med.* **359**, 2121–2134 (2008).

91. Barcia, J. A. *et al.* Deep brain stimulation for obsessive-compulsive disorder: Is the side relevant? *Stereotact. Funct. Neurosurg.* **92**, 31–36 (2014).

92. Zhou, D. D., Wang, W., Wang, G. M., Li, D. Q. & Kuang, L. An updated meta-analysis: Short-term therapeutic effects of repeated transcranial magnetic stimulation in treating obsessive-compulsive disorder. *Journal of Affective Disorders* (2017). doi:10.1016/j.jad.2017.03.033

93. Rehn, S., Eslick, G. D. & Brakoulias, V. A Meta-Analysis of the Effectiveness of Different Cortical Targets Used in Repetitive Transcranial Magnetic Stimulation (rTMS) for the Treatment of Obsessive-Compulsive Disorder (OCD). *Psychiatric Quarterly* (2018). doi:10.1007/s11126-018-9566-7

94. Carmi, L. *et al.* Clinical and electrophysiological outcomes of deep TMS over the medial prefrontal and anterior cingulate cortices in OCD patients. *Brain Stimul.* (2018). doi:10.1016/j.brs.2017.09.004

95. Carmi, L. *et al.* Efficacy and Safety of Deep Transcranial Magnetic Stimulation for Obsessive-Compulsive Disorder: A Prospective Multicenter Randomized Double-Blind Placebo-Controlled Trial. *Am. J. Psychiatry* (2019). doi:10.1176/appi.ajp.2019.18101180

96. Fineberg, N. A. *et al.* Probing compulsive and impulsive behaviors, from animal models to endophenotypes: A narrative review. *Neuropsychopharmacology* (2010). doi:10.1038/npp.2009.185

97. Chamberlain, S. R. *et al.* A neuropsychological comparison of obsessive-compulsive disorder and trichotillomania. *Neuropsychologia* (2007). doi:10.1016/j.neuropsychologia.2006.07.016

98. Bohne, A., Savage, C. R., Deckersbach, T., Keuthen, N. J. & Wilhelm, S. Motor inhibition in trichotillomania and obsessive-compulsive disorder. *J. Psychiatr. Res.* (2008). doi:10.1016/j.jpsychires.2006.11.008

99. Cavedini, P., Zorzi, C., Piccinni, M., Cavallini, M. C. & Bellodi, L. Executive Dysfunctions in Obsessive-Compulsive Patients and Unaffected Relatives: Searching for a New Intermediate Phenotype. *Biol. Psychiatry* (2010). doi:10.1016/j.biopsych.2010.02.012

100. Purcell, R., Maruff, P., Kyrios, M. & Pantelis, C. Neuropsychological deficits in obsessive-compulsive disorder: A comparison with unipolar depression, panic disorder, and normal controls. *Arch. Gen. Psychiatry* (1998). doi:10.1001/archpsyc.55.5.415

101. Lawrence, N. S. *et al.* Decision making and set shifting impairments are associated with distinct symptom dimensions in obsessive-compulsive disorder.

Neuropsychology (2006). doi:10.1037/0894-4105.20.4.409

102. Watkins, L. H. *et al.* Executive function in Tourette's syndrome and obsessive-compulsive disorder. *Psychol. Med.* (2005). doi:10.1017/S0033291704003691

103. John, B. & Lewis, K. R. Chromosome variability and geographic distribution in insects. *Science (80-.).* (1966). doi:10.1126/science.152.3723.711

104. Gottesman, I. I. & Gould, T. D. The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry* (2003). doi:10.1176/appi.ajp.160.4.636

105. Chamberlain, S. R. *et al.* Impaired cognitive flexibility and motor inhibition in unaffected first-degree relatives of patients with obsessive-compulsive disorder. *Am. J. Psychiatry* (2007). doi:10.1176/ajp.2007.164.2.335

106. Rajender, G. *et al.* Study of neurocognitive endophenotypes in drug-naïve obsessive-compulsive disorder patients, their first-degree relatives and healthy controls. *Acta Psychiatr. Scand.* (2011). doi:10.1111/j.1600-0447.2011.01733.x

107. Falkenstein, M., Hohnsbein, J., Hoormann, J. & Blanke, L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalogr. Clin. Neurophysiol.* (1991). doi:10.1016/0013-4694(91)90062-9

108. Gehring, W. J., Himle, J. & Nisenson, L. G. Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychol. Sci.* (2000). doi:10.1111/1467-9280.00206

109. Delorme, R. *et al.* Shared executive dysfunctions in unaffected relatives of patients with autism and obsessive-compulsive disorder. *Eur. Psychiatry* (2007). doi:10.1016/j.eurpsy.2006.05.002

110. Bienvenu, O. J. *et al.* Is obsessive-compulsive disorder an anxiety disorder, and what, if any, are spectrum conditions? A family study perspective. *Psychol. Med.* (2012). doi:10.1017/S0033291711000742

111. L., M. *et al.* Neurocognitive endophenotypes of obsessive-compulsive disorder. *Brain* (2007). doi:http://dx.doi.org/10.1093/brain/awm205

112. Chamberlain, S. R. *et al.* Orbitofrontal dysfunction in patients with obsessive-compulsive disorder and their unaffected relatives. *Science (80-.).* (2008). doi:10.1126/science.1154433

113. Gillan, C. M. *et al.* Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *Am. J. Psychiatry* **168**, 718–726 (2011).

114. Thorndike, E. L. Animal intelligence: An experimental study of the associative processes in animals. *Psychol. Rev.* **2**, 1–107 (1898).

115. Thorndike, E. L. & Jelliffe. Animal Intelligence. Experimental Studies. *The Journal of Nervous and Mental Disease* **39**, 357 (1911).

116. Hull, C. L. *Principles of Behavior: An Introduction to Behavior Theory.* (D. Appleton-Century Company, 1943). doi:10.1037/h0051597

117. Skinner, B. F. *The Behavior of Organisms An Experimental Analysis*. *Journal of General Psychology* (1938). doi:10.1080/00221309.1936.9713156

118. Skinner, B. F. Reinforcement today. *Am. Psychol.* (1958). doi:10.1037/h0049039

119. Skinner, B. F. Operant behavior. *Am. Psychol.* (1963). doi:10.1037/h0045185

120. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).

121. Blodgett, H. C. The effect of the introduction of reward upon the maze performance of rats. *Univ. Calif. Publ. Psychol.* (1929).

122. Tolman, E. & Honzik, C. Introduction and removal of reward, and maze performance in rats. *Univ. Calif. Publ. Psychol.* (1930).

123. Adams, C. D. & Dickinson, A. Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. Sect. B Comp. Physiol. Psychol.* **33**, 109–121 (1981).

124. Adams, C. D. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol. Sect. B* **34**, 77–98 (1982).

125. Dickinson, A. Actions and Habits: The Development of Behavioural Autonomy. *Philos. Trans. R. Soc. B Biol. Sci.* **308**, 67–78 (1985).

126. Colwill, R. M. & Rescorla, R. A. Postconditioning Devaluation of a Reinforcer Affects Instrumental Responding. *J. Exp. Psychol. Anim. Behav. Process.* **11**, 120–132 (1985).

127. Colwill, R. M. & Rescorla, R. A. Instrumental Responding Remains Sensitive to Reinforcer Devaluation After Extensive Training. *J. Exp. Psychol. Anim. Behav. Process.* (1985). doi:10.1037/0097-7403.11.4.520

128. Balleine, B. W. & O'Doherty, J. P. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69 (2010).

129. Robbins, T. W. & Costa, R. M. Habits. *Curr. Biol.* **27**, R1200–R1206 (2017).

130. Dickinson, A., Nicholas, D. J. & Adams, C. D. The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *Q. J. Exp. Psychol. Sect. B* **35**, 35–51 (1983).

131. Dickinson, A. *Contemporary Animal Learning Theory*. *The American Journal of Psychology* (Cambridge University Press, 1980). doi:10.2307/1422669

132. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).

133. Lingawi, N. W., Dezfouli, A. & Balleine, B. W. The Psychological and Physiological Mechanisms of Habit Formation. in *The Wiley Handbook on the Cognitive Neuroscience of Learning* (2015). doi:10.1002/9781118650813.ch16

134. Daw, N. D. & O'Doherty, J. P. Multiple Systems for Value Learning. in *Neuroeconomics: Decision Making and the Brain: Second Edition* (2013).

doi:10.1016/B978-0-12-416008-8.00021-8

135. Gremel, C. M. & Costa, R. M. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat. Commun.* **4**, 2264 (2013).

136. Pavlov, I. P. An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences* (1927). doi:10.5214/ans.0972-7531.1017309

137. Konorski, J. *Integrative activity of the brain: An interdisciplinary approach*. (University of Chicago).

138. Holland, P. C. & Rescorla, R. A. The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning. *J. Exp. Psychol. Anim. Behav. Process.* (1975). doi:10.1037/0097-7403.1.4.355

139. RESCORLA, R. A. Probability of shock in the presence and absence of CS in fear conditioning. *J. Comp. Physiol. Psychol.* (1968). doi:10.1037/h0025984

140. Dickinson, A. & Charnock, D. J. Contingency effects with maintained instrumental reinforcement. *Q. J. Exp. Psychol. Sect. B* **37**, 397–416 (1985).

141. Yin, H. H., Ostlund, S. B. & Balleine, B. W. Reward-guided learning beyond dopamine in the nucleus accumbens: The integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience* (2008). doi:10.1111/j.1460-9568.2008.06422.x

142. Balleine, B. W., Liljeholm, M. & Ostlund, S. B. The integrative function of the basal ganglia in instrumental conditioning. *Behavioural Brain Research* (2009). doi:10.1016/j.bbr.2008.10.034

143. Costa, R. M. A selectionist account of de novo action learning. *Current Opinion in Neurobiology* (2011). doi:10.1016/j.conb.2011.05.004

144. Graybiel, A. M. Building action repertoires: memory and learning functions of the basal ganglia. *Curr. Opin. Neurobiol.* (1995). doi:10.1016/0959-4388(95)80100-6

145. Hikosaka, O. Neural systems for control of voluntary action - A hypothesis. *Advances in Biophysics* (1998). doi:10.1016/S0065-227X(98)80004-X

146. Wickens, J. R., Reynolds, J. N. J. & Hyland, B. I. Neural mechanisms of reward-related motor learning. *Current Opinion in Neurobiology* (2003). doi:10.1016/j.conb.2003.10.013

147. Yin, H. H. & Knowlton, B. J. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* (2006). doi:10.1038/nrn1919

148. Haber, S. N. The primate basal ganglia: Parallel and integrative networks. in *Journal of Chemical Neuroanatomy* (2003). doi:10.1016/j.jchemneu.2003.10.003

149. Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W. & Pennartz, C. M. A. Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neurosciences* (2004). doi:10.1016/j.tins.2004.06.006

150. Yin, H. H., Ostlund, S. B., Knowlton, B. J. & Balleine, B. W. The role of the

dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* **22**, 513–523 (2005).

151. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur. J. Neurosci.* (2005). doi:10.1111/j.1460-9568.2005.04219.x

152. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. in *Neuropharmacology* **37**, 407–419 (1998).

153. Corbit, L. H. & Balleine, B. W. The role of prelimbic cortex in instrumental conditioning. *Behav. Brain Res.* **146**, 145–157 (2003).

154. Ostlund, S. B. & Balleine, B. W. Lesions of Medial Prefrontal Cortex Disrupt the Acquisition But Not the Expression of Goal-Directed Learning. *J. Neurosci.* (2005). doi:10.1523/jneurosci.1921-05.2005

155. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189 (2004).

156. Killcross, S. & Coutureau, E. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* (2003). doi:10.1093/cercor/13.4.400

157. Thorn, C. A., Atallah, H., Howe, M. & Graybiel, A. M. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. *Neuron* (2010). doi:10.1016/j.neuron.2010.04.036

158. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).

159. Hilario, M., Holloway, T., Jin, X. & Costa, R. M. Different dorsal striatum circuits mediate action discrimination and action generalization. *Eur. J. Neurosci.* (2012). doi:10.1111/j.1460-9568.2012.08073.x

160. De Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F. & Dickinson, A. Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *J. Exp. Psychol. Anim. Behav. Process.* (2007). doi:10.1037/0097-7403.33.1.1

161. de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A. & Fletcher, P. C. Differential Engagement of the Ventromedial Prefrontal Cortex by Goal-Directed and Habitual Behavior toward Food Pictures in Humans. *J. Neurosci.* (2009). doi:10.1523/jneurosci.1639-09.2009

162. Valentin, V. V., Dickinson, A. & O'Doherty, J. P. Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci.* **27**, 4019–26 (2007).

163. Tanaka, S. C., Balleine, B. W. & O'Doherty, J. P. Calculating Consequences: Brain Systems That Encode the Causal Effects of Actions. *J. Neurosci.* **28**, 6750–6755 (2008).

164. Tricomi, E., Balleine, B. W. & O'Doherty, J. P. A specific role for posterior

dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* **29**, 2225–2232 (2009).

165. de Wit, S. *et al.* Shifting the balance between goals and habits: Five failures in experimental habit induction. *J. Exp. Psychol. Gen.* (2018). doi:10.1037/xge0000402

166. Dias-Ferreira, E. *et al.* Chronic stress causes frontostriatal reorganization and affects decision-making. *Science (80-.).* (2009). doi:10.1126/science.1171203

167. Sutton, R. S. & Barto, A. G. Introduction to Reinforcement Learning. **4**, (1998).

168. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* (1996).

169. Doya, K. Reinforcement learning: Computational theory and biological mechanisms. *HFSP J.* (2007). doi:10.2976/1.2732246/10.2976/1

170. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–11 (2005).

171. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction (2nd edition)*. (MIT Press, 2018).

172. Thorndike, E. L. Animal intelligence: An experimental study of the associative processes in animals. *Psychol. Rev.* **2**, 1–107 (1898).

173. Sutton, R. S. Learning to Predict by the Methods of Temporal Differences. *Mach. Learn.* (1988). doi:10.1023/A:1022633531479

174. Niv, Y., Joel, D. & Dayan, P. A normative perspective on motivation. *Trends Cogn. Sci.* (2006). doi:10.1016/j.tics.2006.06.010

175. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (80-.).* (1997). doi:10.1126/science.275.5306.1593

176. Berns, G. S., McClure, S. M., Pagnoni, G. & Montague, P. R. Predictability Modulates Human Brain Response to Reward. *J. Neurosci.* (2001). doi:10.1523/jneurosci.21-08-02793.2001

177. O’Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* (2003). doi:10.1016/S0896-6273(03)00169-7

178. Haruno, M. A Neural Correlate of Reward-Based Behavioral Learning in Caudate Nucleus: A Functional Magnetic Resonance Imaging Study of a Stochastic Decision Task. *J. Neurosci.* (2004). doi:10.1523/jneurosci.3417-03.2004

179. D&apos;Ardenne, K., McClure, S. M., Nystrom, L. E. & Cohen, J. D. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science (80-.).* (2008). doi:10.1126/science.1150605

180. Gläscher, J., Daw, N., Dayan, P. & O’Doherty, J. P. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).

181. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).

182. Wunderlich, K., Dayan, P. & Dolan, R. J. Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* **15**, 786–791 (2012).

183. Wan Lee, S., Shimojo, S. & O'Doherty, J. P. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron* **81**, 687–699 (2014).

184. Voon, V. *et al.* Disorders of compulsivity: a common bias towards learning habits. *Mol. Psychiatry* **20**, 345–352 (2015).

185. Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M. & Barch, D. M. Reduced model-based decision-making in schizophrenia. *J. Abnorm. Psychol.* **125**, 777–787 (2016).

186. Sebold, M. *et al.* Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology* **70**, 122–131 (2014).

187. Eppinger, B., Walter, M., Heekeren, H. R. & Li, S. C. Of goals and habits: Age-related and individual differences in goal-directed decision-making. *Front. Neurosci.* (2013). doi:10.3389/fnins.2013.00253

188. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci.* **110**, 20941–20946 (2013).

189. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* (2013). doi:10.1177/0956797612463080

190. Schad, D. J. *et al.* Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Front. Psychol.* **5**, (2014).

191. Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D. & Dolan, R. J. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron* **80**, 914–919 (2013).

192. Friedel, E. *et al.* Devaluation and sequential decisions: linking goal-directed and model-based behavior. *Front. Hum. Neurosci.* **8**, (2014).

193. Kaufman, Arnold; Baron, Alan; and Kopp, R. E. Some Effects of Instructions on Human Operant Behavior. *Psychon. Monogr. Suppl.* **1**, 243–50 (1966).

194. Baron, A., Kaufman, A. & Stauber, K. A. Effects of instructions and reinforcement-feedback on human operant behavior maintained by fixed-interval reinforcement1. *J. Exp. Anal. Behav.* (1969). doi:10.1901/jeab.1969.12-701

195. Baron, A. & Galizio, M. Instructional control of human operant behavior. *Psychol. Rec.* (1983).

196. Wilson, G. D. Reversal of Differential GSR Conditioning by Instructions. *J. Exp. Psychol.* **76**, 491–93 (1968).

197. Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D. & Phelps, E. A. Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *Elife* (2016). doi:10.7554/elife.15192

198. Doll, B. B., Jacobs, W. J., Sanfey, A. G. & Frank, M. J. Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res.* **1299**, 74–94 (2009).

199. Biele, G., Rieskamp, J. & Gonzalez, R. Computational models for the combination of advice and individual learning. *Cogn. Sci.* (2009). doi:10.1111/j.1551-6709.2009.01010.x

200. Li, J., Delgado, M. R. & Phelps, E. A. How instructed knowledge modulates the neural systems of reward learning. *Proc. Natl. Acad. Sci.* (2011). doi:10.1073/pnas.1014938108

201. Benito, K. & Storch, E. A. Assessment of obsessive–compulsive disorder: review and future directions. *Expert Rev. Neurother.* **11**, 287–298 (2011).

202. Goodman, W. K. *et al.* The Yale-Brown Obsessive Compulsive Scale: II. Validity. *Arch. Gen. Psychiatry* **46**, 1012–1016 (1989).

203. Storch, E. A., Lewin, A. B., De Nadai, A. S. & Murphy, T. K. Defining treatment response and remission in obsessive-compulsive disorder: A signal detection analysis of the children's yale-brown obsessive compulsive scale. *J. Am. Acad. Child Adolesc. Psychiatry* **49**, 708–717 (2010).

204. Lewin, A. B. *et al.* Refining clinical judgment of treatment outcome in obsessive-compulsive disorder. *Psychiatry Res.* **185**, 394–401 (2011).

205. Simpson, H. B., Huppert, J. D., Petkova, E., Foa, E. B. & Liebowitz, M. R. Response versus remission in obsessive-compulsive disorder. *J. Clin. Psychiatry* **67**, 269–276 (2006).

206. Farris, S. G., McLean, C. P., Van Meter, P. E., Simpson, H. B. & Foa, E. B. Treatment response, symptom remission, and wellness in obsessive-compulsive disorder. *J. Clin. Psychiatry* **74**, 685–690 (2013).

207. Nestadt, G. *et al.* The relationship between obsessive-compulsive disorder and anxiety and affective disorders: results from the Johns Hopkins OCD Family Study. *Psychol. Med.* **31**, 481–7 (2001).

208. Moritz, S. *et al.* Dimensional structure of the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS). *Psychiatry Res.* **109**, 193–199 (2002).

209. Rapp, A. M., Bergman, R. L., Piacentini, J. & McGuire, J. F. Evidence-Based Assessment of Obsessive-Compulsive Disorder. *J. Cent. Nerv. Syst. Dis.* **8**, 13–29 (2016).

210. Stein, D. J. Obsessive-compulsive disorder. in *Lancet* **360**, 397–405 (2002).

211. First, M. B. et, Spitzer, R. L., Gibbon, M. & Williams, J. B. W. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient*

Edition. *for DSMIV* (2002).

212. Del-Ben, C. M. *et al.* Reliability of the Structured Clinical Interview for DSM-IV--Clinical Version translated into Portuguese. [Portuguese]. [References]. *Rev. Bras. Psiquiatr.* **23**, 159 (2001).

213. Sheehan, D. V. *et al.* The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *Eur. Psychiatry* **12**, 232–241 (1997).

214. Guterres, T., Levy, P. & Amorim, P. MINI International Neuropsychiatric Interview: Versão Português (Portugal) 5.0.0. (1999).

215. Beck, A. T., Steer, R. A. & Brown, G. K. Manual for the Beck depression inventory-II. *San Antonio, TX Psychol. Corp.* 1–82 (1996).

216. Campos, R. & Gonçalves, B. The Portuguese Version of the Beck Depression Inventory-II (BDI-II). *Eur. J. Psychol. Assess.* **27**, 258–264 (2011).

217. Spielberger, C. Manual for the State-Trait Anxiety Inventory (STAI). *Consult. Psychol. Press* 4–26 (1983).

218. Barnes, L. L. B., Harp, D. & Jung, W. S. Reliability generalization of scores on the spielberger state-trait anxiety inventory. *Educ. Psychol. Meas.* **62**, 603–618 (2002).

219. Santos, S. & Silva, D. Adaptação do State-Trait Anxiety Inventory (STAI)- Forma Y para a população portuguesa: Primeiros dados. *Rev. Port. Psicol.* **32**, 85–98 (1997).

220. Galhardo, A. & Pinto-Gouveia, J. Inventário Obsessivo de Coimbra: Avaliação de Obsessões e Compulsões. *Psychologica* **48**, 121–124 (2008).

221. Furr, R. M. Confirmatory factor analysis. in *Scale Construction and Psychometrics for Social and Personality Psychology* 1–30 (2011). doi:10.4135/9781446287866

222. Frost, R. O., Steketee, G., Krause, M. S. & Trepanier, K. L. The Relationship of the Yale-Brown Obsessive Compulsive Scale (YBOCS) to Other Measures of Obsessive Compulsive Symptoms in a Nonclinical Population. *J. Pers. Assess.* **65**, 158–168 (1995).

223. Jin, P. H. *et al.* Clinical correlates of recurrent major depression in obsessive-compulsive disorder. *Depression and Anxiety* **20**, 86–91 (2004).

224. Kolada, J. L., Bland, R. C. & Newman, S. C. Obsessive-Compulsive Disorder. *Acta Psychiatr. Scand.* **89**, 24–35 (1994).

225. Rasmussen, S. A. & Tsuang, M. T. Clinical Characteristics and Family History in DSM-III Obsessive-Compulsive Disorder. *Am. J. Psychiatry* **143**, 317–322 (1986).

226. Sloman, S. A. The empirical case for two systems of reasoning. *Psychol. Bull.* **119**, 3–22 (1996).

227. Kahneman, D. A perspective on judgment and choice : Mapping bounded rationality . By Kahneman , Daniel American Psychologist . 2003 Sep Vol 58 (9

) 697-720. *Behav. Sci.* **58**, 697–720 (2003).

228. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).

229. Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M. & Dolan, R. J. Model-Based Reasoning in Humans Becomes Automatic with Training. *PLoS Comput. Biol.* **11**, (2015).

230. Worbe, Y. *et al.* Valence-dependent influence of serotonin depletion on model-based choice strategy. *Mol. Psychiatry* **21**, 624–629 (2016).

231. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychol. Sci.* **24**, 751–761 (2013).

232. Skatova, A., Chan, P. A. & Daw, N. D. Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Front. Hum. Neurosci.* **7**, (2013).

233. Radenbach, C. *et al.* The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology* **53**, 268–280 (2015).

234. Deserno, L. *et al.* Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc. Natl. Acad. Sci.* **112**, 1595–1600 (2015).

235. Voon, V. *et al.* Disorders of compulsivity: a common bias towards learning habits. *Mol. Psychiatry* **20**, 345–352 (2015).

236. Voon, V. *et al.* Motivation and value influences in the relative balance of goal-directed and habitual behaviours in obsessive-compulsive disorder. *Transl. Psychiatry* **5**, e670 (2015).

237. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. *Elife* **5**, (2016).

238. Akam, T., Costa, R. & Dayan, P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Comput. Biol.* **11**, (2015).

239. Bostan, A. C. & Strick, P. L. The basal ganglia and the cerebellum: nodes in an integrated network. *Nature Reviews Neuroscience* 1–13 (2018). doi:10.1038/s41583-018-0002-7

240. Biele, G., Rieskamp, J., Krugel, L. K. & Heekeren, H. R. The Neural basis of following advice. *PLoS Biol.* (2011). doi:10.1371/journal.pbio.1001089

241. Kool, W., Cushman, F. A. & Gershman, S. J. When Does Model-Based Control Pay Off? *PLoS Comput. Biol.* **12**, (2016).

242. Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).

243. Goodman, W. K. *et al.* The Yale-Brown Obsessive. *Arch. Gen. Psychiatry* **46**, 1006–1011 (1989).

244. Berch, D. B., Krikorian, R. & Huha, E. M. The Corsi block-tapping task: methodological and theoretical considerations. *Brain Cogn.* **38**, 317–38 (1998).

245. Lovibond, S. H. & Lovibond, P. F. *Manual for the Depression Anxiety Stress Scales*. *Psychology Foundation of Australia* (1995). doi:DOI: 10.1016/0005-7967(94)00075-U

246. Huys, Q. J. M. *et al.* Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput. Biol.* **7**, (2011).

247. McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y. & Schoenbaum, G. Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. *J. Neurosci.* (2011). doi:10.1523/jneurosci.5499-10.2011

248. Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science (80-.).* (2012). doi:10.1126/science.1227489

249. Corbit, L. H., Muir, J. L. & Balleine, B. W. Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *Eur. J. Neurosci.* (2003). doi:10.1046/j.1460-9568.2003.02833.x

250. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behav. Brain Res.* **166**, 189–196 (2006).

251. Coutureau, E. & Killcross, S. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behav. Brain Res.* **146**, 167–174 (2003).

252. Reber, J. *et al.* Selective impairment of goal-directed decision-making following lesions to the human ventromedial prefrontal cortex. *Brain* (2017). doi:10.1093/brain/awx105

253. Gremel, C. M. & Costa, R. M. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat. Commun.* **4**, (2013).

254. Cohen, J. D. *et al.* Computational approaches to fMRI analysis. *Nature Neuroscience* (2017). doi:10.1038/nn.4499

255. Wimmer, G. E., Daw, N. D. & Shohamy, D. Generalization of value in reinforcement learning by humans. *Eur. J. Neurosci.* (2012). doi:10.1111/j.1460-9568.2012.08017.x

256. Elliott, R. Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies. *Cereb. Cortex* (2000). doi:10.1093/cercor/10.3.308

257. Rolls, E. T. The Orbitofrontal Cortex and Reward. *Cereb. Cortex* (2000). doi:10.1093/cercor/10.3.284

258. Kringelbach, M. L. & Rolls, E. T. The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology* (2004). doi:10.1016/j.pneurobio.2004.03.006

259. Noonan, M. P. *et al.* Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* (2010). doi:10.1073/pnas.1012246107

260. Rotge, J. Y. *et al.* Anatomical Alterations and Symptom-Related Functional Activity in Obsessive-Compulsive Disorder Are Correlated in the Lateral Orbitofrontal Cortex. *Biological Psychiatry* (2010). doi:10.1016/j.biopsych.2009.10.007

261. Remijnse, P. L. *et al.* Reduced orbitofrontal-striatal activity on a reversal learning task in obsessive-compulsive disorder. *Arch. Gen. Psychiatry* (2006). doi:10.1001/archpsyc.63.11.1225

262. Harrison, B. J. *et al.* Altered corticostriatal functional connectivity in obsessive-compulsive disorder. *Arch. Gen. Psychiatry* (2009). doi:10.1001/archgenpsychiatry.2009.152

263. Banca, P. *et al.* Imbalance in habitual versus goal directed neural systems during symptom provocation in obsessive-compulsive disorder. *Brain* **138**, 798–811 (2015).

264. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* (2009). doi:10.1016/j.neuroimage.2009.06.060

265. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* (2001). doi:10.1016/S1361-8415(01)00036-6

266. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* (2002).

267. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* (2002). doi:10.1002/hbm.10062

268. Woolrich, M. W., Ripley, B. D., Brady, M. & Smith, S. M. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* (2001). doi:10.1006/nimg.2001.0931

269. Beckmann, C. F., Jenkinson, M. & Smith, S. M. General multilevel linear modeling for group analysis in FMRI. *Neuroimage* (2003). doi:10.1016/S1053-8119(03)00435-X

270. Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M. & Smith, S. M. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage* (2004). doi:10.1016/j.neuroimage.2003.12.023

271. Woolrich, M. Robust group analysis using outlier inference. *Neuroimage* (2008). doi:10.1016/j.neuroimage.2008.02.042

272. Jasper, H. & Penfield, W. Electrocorticograms in man: Effect of voluntary

movement upon the electrical activity of the precentral gyrus. *Arch. Psychiatr. Nervenkr.* (1949). doi:10.1007/BF01062488

273. Yousry, T. A. *et al.* Localization of the motor hand area to a knob on the precentral gyrus. A new landmark. *Brain* (1997). doi:10.1093/brain/120.1.141

274. Knutson, B., Adams, C. M., Fong, G. W. & Hommer, D. Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens. *J. Neurosci.* (2001). doi:10.1523/jneurosci.21-16-j0002.2001

275. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* (1997).

276. Schwarzlose, R. F., Baker, C. I. & Kanwisher, N. Separate face and body selectivity on the fusiform gyrus. *J. Neurosci.* (2005). doi:10.1523/JNEUROSCI.2621-05.2005

277. Peelen, M. V. & Downing, P. E. Selectivity for the human body in the fusiform gyrus. *J. Neurophysiol.* (2005). doi:10.1152/jn.00513.2004

278. Diedrichsen, J. & Kriegeskorte, N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* (2017). doi:10.1371/journal.pcbi.1005508

279. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* (2008). doi:10.3389/neuro.06.004.2008

280. Nili, H. *et al.* A Toolbox for Representational Similarity Analysis. *PLoS Comput. Biol.* (2014). doi:10.1371/journal.pcbi.1003553

281. Taylor, S. Assessment of obsessions and compulsions: Reliability, validity, and sensitivity to treatment effects. *Clin. Psychol. Rev.* (1995). doi:10.1016/0272-7358(95)00015-H

282. Guy, W. CGI Clinical Global Impressions. *ECDEU Assess. Man.* (1976).

283. Fernandes, B. S. *et al.* The new field of 'precision psychiatry'. *BMC Med.* (2017). doi:10.1186/s12916-017-0849-x

284. Singh, I. & Rose, N. Biomarkers in psychiatry. *Nature* (2009). doi:10.1038/460202a

285. McGorry, P. *et al.* Biomarkers and clinical staging in psychiatry. *World Psychiatry* (2014). doi:10.1002/wps.20144

286. Venkatasubramanian, G. & Keshavan, M. S. Biomarkers in psychiatry – A critique. *Annals of Neurosciences* (2016). doi:10.1159/000443549

287. Kalia, M. & Costa E Silva, J. Biomarkers of psychiatric diseases: Current status and future prospects. *Metabolism.* (2015). doi:10.1016/j.metabol.2014.10.026

288. Yahata, N., Kasai, K. & Kawato, M. Computational neuroscience approach to biomarkers and treatments for mental disorders. *Psychiatry and Clinical Neurosciences* (2017). doi:10.1111/pcn.12502

289. Montague, P. R., Dolan, R. J. & Friston, K. J. Computational psychiatry. *Trends Cogn. …* (2012). doi:10.1016/j.tics.2011.11.018

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

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

292. Maia, T. V., Huys, Q. J. M. & Frank, M. J. Theory-Based Computational Psychiatry. *Biological Psychiatry* (2017). doi:10.1016/j.biopsych.2017.07.016

293. A M Turing, I. B. Computing machinery and intelligence. *Mind* (1950).

294. Collins, A. G. E. Reinforcement learning: bringing together computation and cognition. *Current Opinion in Behavioral Sciences* (2019). doi:10.1016/j.cobeha.2019.04.011

295. Keramati, M., Dezfouli, A. & Piray, P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* (2011). doi:10.1371/journal.pcbi.1002055

296. Pezzulo, G., Rigoli, F. & Chersi, F. The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Front. Psychol.* (2013). doi:10.3389/fpsyg.2013.00092

297. van der Meer, M. A. A. Covert expectation-of-reward in rat ventral striatum at decision points. *Front. Integr. Neurosci.* (2009). doi:10.3389/neuro.07.001.2009

298. Shadmehr, R. Control of movements and temporal discounting of reward. *Current Opinion in Neurobiology* (2010). doi:10.1016/j.conb.2010.08.017

299. Fermin, A. S. R. *et al.* Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Sci. Rep.* (2016). doi:10.1038/srep31378

300. Shahar, N. *et al.* Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proc. Natl. Acad. Sci.* (2019). doi:10.1073/pnas.1821647116

301. Ayllon, T. & Azrin, N. H. Reinforcement and instructions with mental patients. *J. Exp. Anal. Behav.* (1964). doi:10.1901/jeab.1964.7-327

302. Galizio, M. Contingency-shaped and rule-governed behavior: instructional control of human loss avoidance. *J. Exp. Anal. Behav.* (1979). doi:10.1901/jeab.1979.31-53

303. Niv, Y. & Langdon, A. Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences* (2016). doi:10.1016/j.cobeha.2016.04.005

304. Baddeley, A. Working Memory. *Science (80-.).* (1992). doi:10.1126/science.1736359

305. Baddeley, A. Working memory: theories, models and controversies. *Annu. Rev. Psychol.* **63**, 1–29 (2012).

306. Collins, A. G. E. The tortoise and the hare: Interactions between reinforcement learning and working memory. *J. Cogn. Neurosci.* (2017). doi:10.1162/jocn_a_01238

307. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* (2012). doi:10.1111/j.1460-9568.2011.07980.x

308. Van Der Wee, N. J. A. *et al.* Spatial working memory deficits in obsessive compulsive disorder are associated with excessive engagement of the medial frontal cortex. *Neuroimage* (2003). doi:10.1016/j.neuroimage.2003.05.001

309. Nakao, T. *et al.* Working memory dysfunction in obsessive-compulsive disorder: A neuropsychological and functional MRI study. *J. Psychiatr. Res.* (2009). doi:10.1016/j.jpsychires.2008.10.013

310. Graybiel, A. M. & Rauch, S. L. Toward a neurobiology of obsessive-compulsive disorder. *Neuron* **28**, 343–347 (2000).

311. Gillan, C. M. *et al.* Enhanced avoidance habits in obsessive-compulsive disorder. *Biol. Psychiatry* **75**, 631–638 (2014).

312. Sjoerds, Z. *et al.* Behavioral and neuroimaging evidence for overreliance on habit learning in alcohol-dependent patients. *Transl. Psychiatry* (2013). doi:10.1038/tp.2013.107

313. Wheaton, M. G., Gillan, C. M. & Simpson, H. B. Does cognitive-behavioral therapy affect goal-directed planning in obsessive-compulsive disorder? *Psychiatry Res.* (2019). doi:10.1016/j.psychres.2018.12.079

314. Kalanthroff, E., Abramovitch, A., Steinman, S. A., Abramowitz, J. S. & Simpson, H. B. The chicken or the egg: What drives OCD? *Journal of Obsessive-Compulsive and Related Disorders* (2016). doi:10.1016/j.jocrd.2016.07.005

315. Karno, M., Golding, J. M., Sorenson, S. B. & Burnam, M. A. The Epidemiology of Obsessive-Compulsive Disorder in Five US Communities. *Arch. Gen. Psychiatry* (1988). doi:10.1001/archpsyc.1988.01800360042006

316. Swedo, S. E., Rapoport, J. L., Leonard, H., Lenane, M. & Cheslow, D. Obsessive-Compulsive Disorder in Children and Adolescents: Clinical Phenomenology of 70 Consecutive Cases. *Arch. Gen. Psychiatry* (1989). doi:10.1001/archpsyc.1989.01810040041007

317. Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S. & Ersche, K. D. Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences* (2012). doi:10.1016/j.tics.2011.11.009

318. Gillan, C. M. & Sahakian, B. J. Which is the driver, the obsessions or the compulsions, in OCD? *Neuropsychopharmacology* (2015). doi:10.1038/npp.2014.201

319. Bernardo Barahona-Corrêa, J., Camacho, M., Castro-Rodrigues, P., Costa, R. & Oliveira-Maia, A. J. From thought to action: How the interplay between neuroscience and phenomenology changed our understanding of obsessive-

compulsive disorder. *Frontiers in Psychology* **6**, (2015).

320. Gillan, C. M. *et al.* Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. *Am. J. Psychiatry* (2015). doi:10.1176/appi.ajp.2014.14040525

321. Wood, J. & Ahmari, S. E. A Framework for Understanding the Emerging Role of Corticolimbic-Ventral Striatal Networks in OCD-Associated Repetitive Behaviors. *Front. Syst. Neurosci.* (2015). doi:10.3389/fnsys.2015.00171

322. Milad, M. R. *et al.* Deficits in conditioned fear extinction in obsessive-compulsive disorder and neurobiological changes in the fear circuit. *JAMA Psychiatry* (2013). doi:10.1001/jamapsychiatry.2013.914

323. McLaughlin, N. C. R. *et al.* Extinction retention and fear renewal in a lifetime obsessive-compulsive disorder sample. *Behav. Brain Res.* (2015). doi:10.1016/j.bbr.2014.11.011

324. Apergis-Schoute, A. M. *et al.* Neural basis of impaired safety signaling in Obsessive Compulsive Disorder. *Proc. Natl. Acad. Sci.* (2017). doi:10.1073/pnas.1609194114

325. Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P. & Robinson, O. J. Modeling Avoidance in Mood and Anxiety Disorders Using Reinforcement Learning. *Biol. Psychiatry* (2017). doi:10.1016/j.biopsych.2017.01.017

326. A., B. *et al.* Effects of acute stress and anxiety induced by the inhalation of air enriched with CO2 on model-based and model-free behaviour. *European Neuropsychopharmacology* (2015).

327. Gillan, C. M. & Robbins, T. W. Experimentally-induced and real-world acute anxiety have no effect in model-based learning. *bioRxiv* (2019).

328. Alvares, G. A., Balleine, B. W. & Guastella, A. J. Impairments in goal-directed actions predict treatment response to cognitive-behavioral therapy in social anxiety disorder. *PLoS One* **9**, (2014).

329. Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M. & Daw, N. D. Excessive deliberation in social anxiety Abstract : *bioRxiv* (2019). doi:10.1101/522433

330. Pizzagalli, D. A., Iosifescu, D., Hallett, L. A., Ratner, K. G. & Fava, M. Reduced hedonic capacity in major depressive disorder: Evidence from a probabilistic reward task. *J. Psychiatr. Res.* (2008). doi:10.1016/j.jpsychires.2008.03.001

331. Robinson, O. J., Cools, R., Carlisi, C. O., Sahakian, B. J. & Drevets, W. C. Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. *Am. J. Psychiatry* (2012). doi:10.1176/appi.ajp.2011.11010137

332. Herzallah, M. M. *et al.* Learning from negative feedback in patients with major depressive disorder is attenuated by SSRI antidepressants. *Front. Integr. Neurosci.* (2013). doi:10.3389/fnint.2013.00067

333. Castro-Rodrigues, P. & Oliveira-Maia, A. J. Exploring the effects of depression and treatment of depression in reinforcement learning. *Frontiers in Integrative Neuroscience* (2013). doi:10.3389/fnint.2013.00072

334. Huys, Q. J. M., Daw, N. D. & Dayan, P. Depression: A Decision-Theoretic Analysis. *Annu. Rev. Neurosci.* **38**, 1–23 (2015).

335. Huys, Q. J. M. *et al.* Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput. Biol.* (2012). doi:10.1371/journal.pcbi.1002410

336. Abler, B., Walter, H., Erk, S., Kammerer, H. & Spitzer, M. Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage* (2006). doi:10.1016/j.neuroimage.2006.01.001

337. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* (2018). doi:10.1038/s41593-018-0147-8

338. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–278 (2014).

5.5. Conclusions