

The Effects of Dopaminergic Modulation on Model-based Reinforcement Learning

Kristian Reveles Jensen

Towards the degree of Master of Science by Research

Department of Psychiatry, Faculty of Clinical Medicine
Linacre College, University of Oxford
Trinity Term 2016

This thesis contains approximately 27.000 words

Abstract

A functional polymorphism (Val158Met) in the catechol-O-methyltransferase (COMT) gene influences cortical dopamine and prefrontal cortex function in humans, both of which are important in learning and decision making. Moreover, dysfunction in the brain's dopamine circuits and mechanisms of decision making is observed during acute stress.

Mammals are thought to use two evaluation strategies in decision making, model-based and model-free reinforcement learning (RL), which have complementary strengths and weaknesses in terms of computational demand and statistical efficiency. Model-based is frontal cortex-dependent and reflects goal-directed behaviour, while model-free relies on the reward prediction errors from the striatum and reflects habits.

The aim of this thesis was to investigate how acute stress influenced RL, whether RL is altered in a mouse model of altered COMT activity, and if COMT-genotype influences the effects of acute stress on RL.

I tested RL in both wild-type mice and transgenic mice carrying the human Met158 allele. They were tested on a novel, multi-step, probabilistic RL task - the rodent two-step task - which can dissociate the influence of model-free and model-based RL on decision making. I used vehicle injections to induce acute stress.

The acute injection stress markedly impaired model-based learning and shifted decision-making towards more model-free in wild-type animals. No COMT-genotype difference on RL was observed when the mice were not stressed. However, Met homozygotes appeared to be less affected by injection stress on both model-based and model-free RL.

Consistent with findings in humans, I show that acute stress shifts mice towards a more model-free pattern of RL. This highlights the potential confounding effects of injection stress in pharmaco-behavioural experiments. Although formal tests for interactive effects of stress were not possible in some cases, I found that that Met homozygotes appear less sensitive than wild-type animals to the negative effects of injection stress. I hypothesise that this is because they have a greater cognitive reserve that, although unutilised at baseline, can step in when faced with increased cognitive demand resulting from environmental factors (in this case, acute injection stress). These findings suggest that subtle COMT genotype differences in behaviour might be amplified under conditions of stress, providing a potential explanation for some of the controversies seen in the human literature linking COMT genotype and cognitive function.

No part of this thesis has been submitted for any other examination.
The data from the sucrose-reinforced wild-type mice presented in chapter 3 was provided by T. Akam. Who also provided analysis code and statistical models. T. Akam also shared unpublished work on the two-step task, that this thesis relies on. The data from the wild-type mice presented in chapter 3 and 4 were produced in collaboration with C. Korn. We conducted a pharmacological study. In included in my thesis is only the data from non-injection and vehicle-control and none of the drug data. A. Huber provided assistance was provided with animal handling, breeding, genotyping, and colony management. A. Huber, who also provided unpublished work on plasma corticosterone measurements in the transgenic COMT mice presented in chapter 5. All other parts of this thesis were completed by K. Jensen

---

Acknowledgements

Many thanks to my dear friends and colleagues who have been on board my boat and helped we row, navigate when the waters were not calm, and land on time.

Thank you, Prof Liz Tunbridge and Dr Mark Walton for your continuous faith in my mice and me.

Thank you, Clio, Anna, and Thomas for all the great exchanges about science, life in science, and life.

Thank you, Angelika, Alessia, and Diana for all the whisky, veggie burgers, and good times.

Tak til mine ferie og læsegæster, for alle brevene og pakkerne, og støtten hjemmefra.



Figure 1: Team work. Artwork by Merlin Porter (merlinporter.wordpress.com)

Funding

My stay has been financially supported by:

- The Oticon Foundation travel scholarship
- The Augustinus Foundation travel scholarship
- The Medical Society in Copenhagen research scholarship
- Kvinderegensen and Dr Elisabeth H. Taagelund's travel scholarship

This research has been financially supported by:

- Oxford University, Department of Psychiatry
- The Royal Society
- The Wellcome Trust
- The Biotechnology and Biological Sciences Research Council

Table of Contents

ABSTRACT **2**
ACKNOWLEDGEMENTS **3**
FUNDING 4
TABLE OF CONTENTS 5
ABBREVIATIONS **8**
1. INTRODUCTION **9**
1.1. DOPAMINE 9
1.1.1. Dopamine biosynthesis and neurotransmission *9*
1.1.2. Dopamine projections *10*
1.2. CATECHOL-O-METHYLTRANSFERASE AND CORTICAL DOPAMINE 11
1.2.1. Gene expression *12*
1.2.2. Localisation *13*
1.2.3. The COMT Val^{158}Met polymorphism *14*
1.2.4. Cortical dopamine and cognitive function *15*
1.2.5. COMT, PFC dopamine and cognitive function *15*
1.2.5.1. COMT and cognition *16*
1.3. REINFORCEMENT LEARNING AND DOPAMINE 19
1.3.1. Model-free RL and the striatum *19*
1.3.2. Model-based RL and the prefrontal cortex *21*
1.3.3. Summary of model-free and model-based RL *22*
1.3.4. The two-step task *24*
1.4. THE ACUTE STRESS RESPONSE, COMT AND RL 25
1.4.1. Acute stress *25*
1.4.1. Acute stress and model-based/model-free RL *26*
1.4.2. COMT and cognitive function under acute stress *27*
1.5. HYPOTHESES 29
2. GENERAL METHODS **30**
2.1. MICE 30
2.1.1. Pilot cohort (SUCROSE 1) *30*
2.1.2. Second cohort (SUCROSE 2) *31*
2.1.3. COMT cohort *31*
2.1.4. Mice on water deprivation (WATER 1 and WATER 2) *32*
2.1.5. Power and sample size *32*
2.2. THE TWO-STEP REINFORCEMENT LEARNING TASK 33
2.2.1. Testing boxes *33*
2.2.2. Task structure *34*
2.2.3. Task training *38*
2.3. ANALYSIS 39
2.3.1. Reaction times and trial rates *39*
2.3.2. Reversal analysis *39*
2.3.3. Multilevel modelling *40*
2.3.4. Logistic regression *41*
2.3.5. Computational Reinforcement Learning Model *43*
2.3.6. Model comparison *48*
2.3.6.1. Method of model comparison *48*
2.3.6.2. Results of model comparison *49*
2.4. PERMUTATION TESTING 52
2.4.1. Permutation testing within subject and cycle *53*

3. FROM WATER TO SUCROSE REINFORCEMENT 55
3.1. INTRODUCTION 55
3.1.1. Changing the protocol from water to sucrose reinforcement 55
3.2. METHODS 56
3.2.1. Training 57
3.2.1.1. Water Cohorts 57
3.2.1.2. Sucrose Cohorts 57
3.2.2. Analysis 58
3.3. RESULTS 61
3.3.1. Task training and trial rate 61
3.3.2. No performance differences between water and sucrose 64
3.3.3. Difference in reinforcement learning 66
3.4. DISCUSSION 68
3.4.1. Trial rate and training time 70
3.4.2. Performance differences 70

4. ACUTE STRESS ON MODEL-BASED/MODEL-FREE RL 73
4.1. INTRODUCTION 73
4.1.1. Changes in behaviour over time 73
4.1.2. Acute stress 76
4.1.2.1. Choice of stress model 78
4.2. INTRODUCTION 81
4.2.1. Subjects and training 81
4.2.2. Behavioural study 81
4.2.3. Analysis 82
4.3. RESULTS 84
4.3.1. Behavioural changes across sessions 84
4.3.2. Behavioural changes within sessions 87
4.3.3. Stress study results 91
4.4. DISCUSSION 95
4.4.1. Changes across sessions 95
4.4.2. Changes within sessions 96
4.4.3. Acute stress 98

5. SEPARATE AND INTERACTIVE EFFECTS OF COMT GENOTYPE AND STRESS ON MODEL-BASED/MODEL-FREE LEARNING 101
5.1. INTRODUCTION 101
5.1.1. The COMT-Met mouse model 101
5.1.2. COMT and stress response 102
5.1.3. Hypotheses 105
5.2. METHODS 106
5.2.1. COMT training 106
5.2.2. Study of baseline effects of COMT genotype on performance 108
5.2.3. The effect of injection stress on COMT-Met and COMT-WT mouse performance 108
5.2.4. Statistics 109
5.3. RESULTS 110
5.3.1. Age, weight and training 110
5.3.2. Baseline effects of COMT genotype on performance 111
5.3.3. The effect of injection stress on COMT mouse performance 114
5.3.3.1. Trial rate and reaction times 114
5.3.3.2. Reversal analysis 115
5.3.3.3. Logistic regression 117
5.3.3.4. RL model 119
5.3.4. Brief summary of findings 120

5.4. DISCUSSION 121
5.4.1. Genotype study 121
5.4.2. Genotype – Stress interaction study 124

6. GENERAL DISCUSSION 131
7. APPENDIX 139
7.1. APPENDIX TO CHAPTER 1 139
7.2. APPENDIX TO CHAPTER 2 139
7.2.1. DEVELOPMENT OF THE COMT MOUSE MODEL 139
7.2.2. COMT GENOTYPING 140
7.2.3. COUNTERBALANCING OF INJECTIONS FOR SUCROSE 2 140
7.2.4. PICTURE OF TESTING BOX WALL 141
7.3. APPENDIX TO CHAPTER 3 141
7.4. APPENDIX TO CHAPTER 4 141
7.4.1. DIET AND SUCROSE CONTRIBUTION DURING TESTING 141
7.4.2. Permutation testing of Water versus Sucrose 143
7.5. APPENDIX TO CHAPTER 5 146
7.5.1. ACROSS SESSION CHANGES 146
7.5.2. WITHIN SESSION CHANGES 148
7.5.2.1. Within-session changes with water-reinforcement 148
3.4.3. Within-session changes with sucrose-reinforcement 150
7.5.3. INJECTION STRESS STUDY 153
7.5.3.1. Reversal plots 153
7.5.3.2. Logistic regression 154
7.5.4. Reinforcement learning model 154
7.6. APPENDIX TO CHAPTER 6 155
7.6.1. COMT BASELINE 155
7.6.1.1. Cumulative trials 155
7.6.1.2. Reversal analysis 156
7.6.1.3. Logistic regression 156
7.6.1.4. Reinforcement learning model 157
7.6.2. COMT STRESS 158
7.6.2.1. Circadian rhythm of plasma CORT 158
7.6.2.2. Trial rate divided by genotype 158
7.6.2.3. 2nd step Reaction time 159
7.6.2.4. Reversal analysis divided by genotype 159
7.6.2.5. Logistic regression – divided by genotype 160
7.6.2.6. RL model analysis - divided by genotype 161

BIBLIOGRAPHY 162

Abbreviations

ACTH	Adrenocorticotropic hormone
BOLD	Blood oxygen level dependent
COMT	Catechol-O-methyltransferase
CORT	Cortisol or corticosterone
CRH	Corticotropin-releasing hormone
DA	Dopamine
DAT	Dopamine Active Transporter
DBH	dopamine-β-hydroxylase
DDC	DOPA-decarboxylase
DHPA	3,4-di-hydroxyphenylacetaldehyde
DOPAC	3,4-dihydroxy-phenylacetic acid
EDS	Extra-Dimensional Set-shifting
fMRI	Functional magnetic resonance imaging
HPA	Hypothalamic–pituitary–adrenal
HVA	Homovanillic acid
i.p.	Intraperitoneal
L-DOPA	3,4-dihydroxyphenylalanine
Leu	Leucine
MAO	Monoamine oxidase
MB	Model-based
Met	Methionine
MF	Model-free
MHPA	3-methoxy-4-hydroxyphenylacet-aldehyde
mPFC	Medial prefrontal cortex
ms	Millisecond
N/A	Not applicable
OFC	Orbitofrontal cortex
pCORT	Plasma cortisol or corticosterone
PET	Positron emission tomography
PFC	Prefrontal cortex
PNMT	Phenylethanolamine-N-methyltransferase
RL	Reinforcement learning
SD	Standard Deviation
SEM	Standard Error of the Mean
SNP	Single nucleotide polymorphism
TH	Tyrosine hydroxylase
VStr	Ventral Striatum
VTA	Ventral tegmental area
WM	Working memory
WT	Wild-type

1. Introduction

The enzyme catechol-O-methyltransferase (COMT) metabolises catecholamines, including dopamine, and much research has focussed on its role in regulating cortical dopamine as well as cognitive function.

This thesis investigates the effect of injection stress and COMT genotype on reinforcement learning.

In this chapter, I will first give an overview of the dopamine systems and the role of COMT in the prefrontal cortex (PFC). I will then introduce the concept of model-based and model-free reinforcement learning, and the task I used to study these learning processes.

Finally, I will briefly, introduce the psychological stress response and how acute stress affects cognition.

1.1. Dopamine

1.1.1. Dopamine biosynthesis and neurotransmission

Dopamine, like the other catecholamines adrenaline and noradrenaline, is synthesised from the dietary amino acid tyrosine (Figure 2). Dopamine is packaged into synaptic vesicles by the vesicular monoamine transporter for release via exocytosis.

Dopamine signals via five G-protein-coupled metabotropic receptors (D_1-D_5) that are grouped into two functional classes: D_1-like receptors, comprised of D_1 and D_5, are positively coupled to the intracellular adenylate cyclase pathway, while the D_2-like receptors, consisting of D_2, D_3, and D_4, are negatively coupled to this pathway, leading to excitatory and inhibitory effects downstream, respectively (Sealfon and Olanow 2000). D_1, followed by D_2, is the most abundant subtype in the brain. All receptor types are present in the corticolimbic regions, while only D_1 and D_2 are expressed in the striatum.

Following its release, dopamine signalling is terminated by reuptake or catabolism. Dopamine

can be removed from the extracellular space by three transporters: the dopamine transporter (DAT), the noradrenaline transporter, or the extraneuronal monoamine transporter (Jonker & Schinkel, 2004; Moron, Brockington, Wise, Rocha, & Hope, 2002). Following reuptake, dopamine can be recycled and packaged into vesicles for rerelease. Alternatively, dopamine can be catabolised by two three-step pathways initiated by catechol-o-methyltransferase (COMT), and monoamine oxidase (MAO, Figure 2).

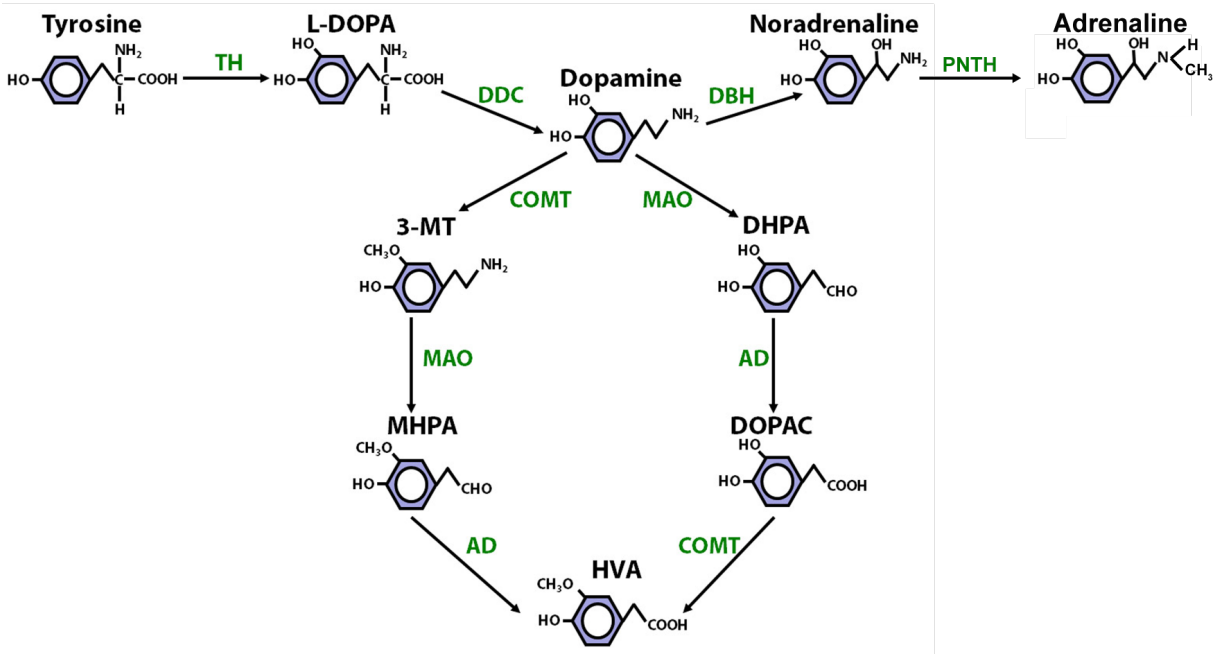


Figure 2: Dopamine synthesis and metabolism. Dopamine is synthesised from the amino acid precursor L-tyrosine. The rate-limiting step is that catalysed by the enzyme tyrosine hydroxylase (TH). The intermediate product 3,4-dihydroxyphenylalanine (L-DOPA) is then converted into dopamine by DOPA-decarboxylase (DDC). In noradrenergic cells, noradrenaline is synthesised from dopamine by dopamine-β-hydroxylase (DBH), and in adrenergic cells noradrenaline is converted into adrenaline by phenylethanolamine-N-methyltransferase (PNMT). Dopamine is degraded by the enzymes catechol-o-methyltransferase (COMT), monoamine oxidase (MAO), and aldehyde dehydrogenase (AD). The intermediate products are 3-methoxytyramine (3-MT), 3,4-di-hydroxyphenylacetaldehyde (DHPA), 3-methoxy-4-hydroxyphenylacet-aldehyde (MHPA), and 3,4-dihydroxy-phenylacetic acid (DOPAC), and the common end product is homovanillic acid (HVA). Image adapted from (Linda M. Laatikainen et al. 2013).

1.1.2. Dopamine projections

Dopamine-synthesising cells in the midbrain project to higher regions of the brain. These can be divided into distinct pathways based on their projections: mesocortical, mesolimbic and nigrostriatal pathways (Björklund and Dunnett) (Table 1).

The mesocortical pathway projects from the ventral tegmental area (VTA) to cortical and limbic structures e.g. the prefrontal cortex (PFC), cingulate cortex, amygdala, and hippocampus, and regulates higher cortical functions (Björklund and Dunnett)(see 1.2.4).

The mesolimbic pathway also originates in the VTA and projects to the core and shell regions of the nucleus accumbens, and the olfactory tubercle, which collectively form the ventral striatum (vStr). Dopamine release in the vStr has been extensively studied in the context of motivation and reinforcement learning (Björklund and Dunnett)(see 1.3.1).

The nigrostriatal pathway originates in the substantia nigra pars compacta and projects mostly to the dorsal striatum. It mediates voluntary motor movement, although emerging findings also implicate it in action selection and decision-making (Hattori 1993).

Most relevant for the studies presented in this thesis are the pathways projecting from the VTA to the PFC and vStr, which are collectively referred to as mesocorticolimbic.

Pathway	Origin	Target	Function
Mesocortical[1]	Ventral tegmental area	Prefrontal cortex, amygdala, hippocampus, brainstem, and other limbic areas	Higher cortical functions
Mesolimbic[1]	Ventral tegmental area	Nucleus accumbens & tuberculum olfactorium (ventral striatum)	Arousal, locomotor activity and reward- related states
Nigrostriatal	Substantia nigra pars compacta	Nucleus caudatus & putamen (dorsal striatum)	Motor control and some cognitive functions

Table 1: Summary of the major dopaminergic pathways in the brain. [1]Due to the mesocortical and mesolimbic pathway having the same origin, it is sometimes referred to collectively as the *meso-corticolimbic* pathway. A fourth pathway from the hypothalamus to the pituitary gland influences the secretion of prolactin (not shown).

1.2. Catechol-O-methyltransferase and cortical dopamine

The O-methylation of catecholamines was observed in 1957 by Nobel Laureate J. Axelrod, who further characterised and named the COMT enzyme (Axelrod 1957; Axelrod and Tomchick 1958). COMT catabolises catechol-containing compounds by adding a methyl group, donated from

S-adenosyl-L-methionine to one of the hydroxyl groups on the catechol-ring of its substrates.

1.2.1. Gene expression

The human COMT gene is localised on chromosome 22, band q11.2 (Grossman et al. 1992) (Figure 2), a region of interest for developmental neuropsychiatric disorders, psychosis and bipolar spectrum disorder (Hiroi et al. 2013). The COMT promoter regions have binding sites and hormone-response elements for oestrogens and glucocorticoids, which regulate COMT expression (Al-Hendy and Salama 2006; Xie et al. 1999). These regulatory sites may contribute to the sexually dimorphic and stress-related effects of COMT e.g. predisposition for psychiatric disorders (C. Chen et al. 2011a; Gurvich and Rossell 2015; Harrison and Tunbridge 2008; Lamb et al. 2016; Francesco Papaleo et al. 2015; E. M. Tunbridge and Harrison 2010; White et al. 2014).

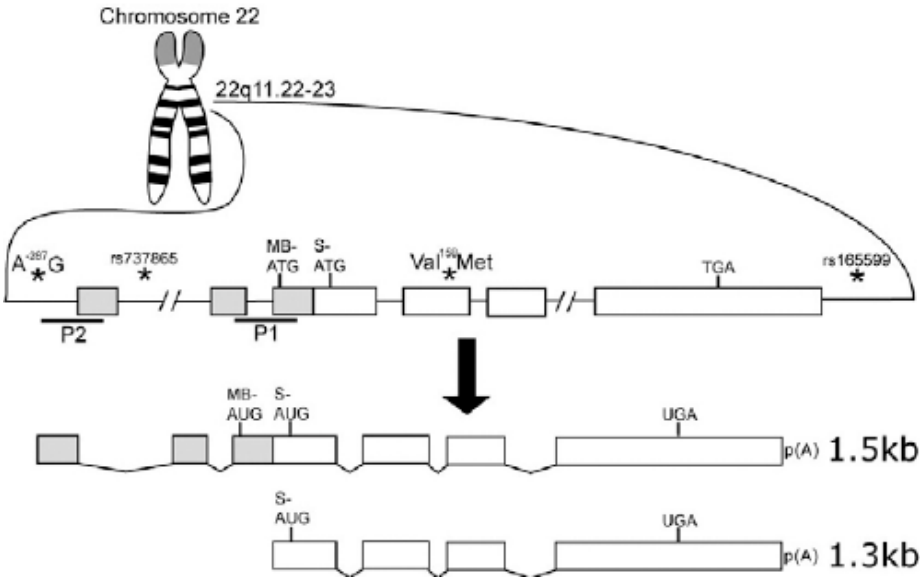


Figure 3: The human COMT gene is localised on chromosome 22, band q11.2 and has six exons (the first two are non-coding). The gene encodes two isoforms from alternative promoters: The P1 promoter guides transcription of the shorter mRNA transcript (1.3 kb), which encodes a soluble isoform of COMT, while the P2 promoter guides transcription of the longer mRNA (1.5 kb), which has two start codons (AUG) and can encode both membrane-bound and soluble COMT isoforms. From (Elizabeth M Tunbridge et al. 2006b)

Membrane-bound COMT is dominant in the human brain (~70%) while soluble COMT is dominant in peripheral tissues, ~70%, (Tenhunen et al. 1994). Membrane-bound COMT also has 10 times greater affinity for dopamine and noradrenaline and is therefore thought to play the

predominant role in the inactivation of catecholamine neurotransmitters, whilst soluble COMT is predominantly responsible for eliminating toxic exogenous catechols in the periphery (Kaakkola et al. 1994; Mannisto et al. 1992).

1.2.2. Localisation

Northern blots show that COMT is present throughout the CNS in humans (Hong et al. 1998; Matsumoto et al. 2003). An *in situ* hybridisation study found COMT highly expressed in the PFC, with lower levels in the striatum, and lowest levels in the VTA and substantia nigra (Matsumoto et al. 2003). In contrast, the dopamine transporter (DAT) is abundant in the striatum, but not in the hippocampus and cortex, has a 1000-fold higher affinity for dopamine than COMT, and faster turn-over (J. Chen et al. 2004). Accordingly, DAT is considered primarily responsible for termination of striatal dopamine signalling, and COMT for cortical dopamine (Figure 4).

The COMT enzyme in expressed in neurons (especially in pyramidal cells in cortical layer II, III, VI) (Matsumoto et al. 2003), in the cell body, in axons and dendrites (J. Chen et al. 2011b), and in postsynaptic spines (Kastner et al. 1994). Other studies suggest, that is also present in glial cells (Myohanen and Mannisto 2010). A series of elaborates experiments identified that membrane-bound COMT is also situated in the pre- and postsynaptic membranes and with the catalytic domain protruding into the extracellular space, whereby COMT can inactivate synaptic and extrasynaptic dopamine (Figure 4, (J. Chen et al. 2011b)).

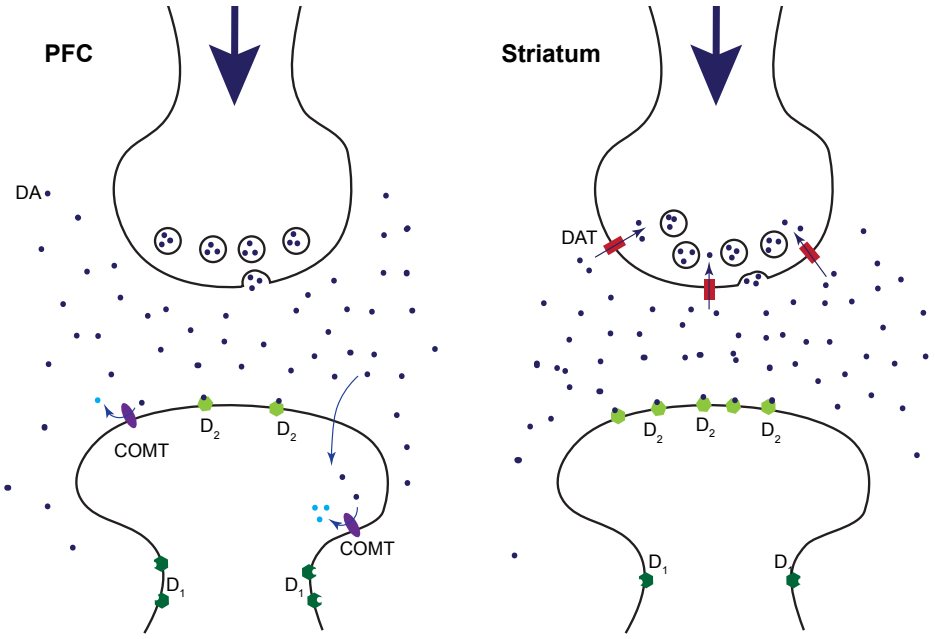


Figure 4: Dopaminergic transmission in the PFC and striatum. In the PFC, the lack of dopamine transporters means that COMT plays a dominant role in terminating dopamine signalling. The lower activity of the COMT-Met results in increased PFC dopamine. In the striatum, the high affinity and abundant DAT is responsible for removing dopamine from the synaptic cleft for recycling into new vesicles. Adapted from (Elizabeth M Tunbridge et al. 2006b).

1.2.3. The COMT Val^{158}Met polymorphism

A functional single nucleotide polymorphism (guanine/adenine [rs4680]) results in a Valine (Val) to Methionine (Met) substitution at position 158 in the membrane-bound form of COMT (Lotta et al. 1995). The Met form is less thermostable and consequently less active at physiological temperatures (J. Chen et al. 2004; Lotta et al. 1995).

The alleles are codominant, and the Val158Met polymorphism is thought to primarily underlie the observed trimodal distribution of COMT activity in human populations (Floderus et al. 1981). The polymorphism is, however, not present in rodents or nonhuman primates, and the high activity Val form of the enzyme is the ancestral form (Palmatier et al. 1999). Rodents have a Leucine (Leu) in the Val/Met equivalent position, though the thermostability and activity of the rodent Leu form is similar to, or slightly higher than, the human Val form (J. Chen et al. 2004; Lotta et al. 1995).

1.2.4. Cortical dopamine and cognitive function

Cortical dopamine is involved in cognition and behaviour (Goldman-Rakic 1998).

Initial work found that catecholamine depletion of the PFC impaired spatial memory in rhesus monkeys (Brozoski et al. 1979) and, via microdialysis, it was possible to observe increases in extracellular dopamine in the PFC during working memory tasks in both primates and rodents (Phillips et al. 2004; Watanabe et al. 1997). However, D_1 receptor agonists impaired, rather than improved, cognitive performance, similar to dopamine depletion (Goldman-Rakic 1998; Sawaguchi and Goldman-Rakic 1991; Williams and Goldman-Rakic 1995). This lead to the concept of an Inverted-U relationship of D_1 receptor activation and working memory; whereby performance is optimal at an intermediate level of receptor activation ((Goldman-Rakic et al. 2000), Figure 6 left). This relationship has been observed in humans, primates and rodents (Mattay et al. 2003; Vijayraghavan et al. 2007).

1.2.5. COMT, PFC dopamine and cognitive function

COMT is highly expressed in the PFC, but relatively lacking in the striatum (Matsumoto et al. 2003). Three lines of evidence suggests that COMT is important for termination of dopamine signalling in the PFC, with degradation by MOA or reuptake by DAT playing a more minor role. Firstly, 3-Methoxytyramine, the first degradation product of dopamine O-methylation by COMT (Figure 2), is the major metabolite of released dopamine in the rat frontal cortex (Karoum et al. 1994). Kaorum et al. (1994) determined that COMT contributed to over 60% of dopamine turnover in the PFC but only 15% in the striatum. Secondly, COMT knockout mice, compared to wild-type controls, had a two-fold increase in dopamine tissue content and a 50% increase in extracellular dopamine levels in the PFC (measured by microdialysis), without an effect in the striatum (Gogos et al. 1998; Käenmäki et al. 2010). Moreover, the elimination of electrically-stimulated phasic

dopamine release was two-fold slower in COMT knockout mice, without any effect in the dorsal striatum (Yavich et al. 2007). Thirdly, whilst DAT is sparse in the PFC, that which is present is located distant from the sites of synaptic release (Lewis et al. 2001; Sesack et al. 1998). In contrast, in the striatum DAT is abundant and localated close to the synapse. It is therefore in an ideal position to limit dopamine spill-over from the synaptic cleft (Hersch et al. 1997). Therefore, DAT is the primary mechanism for terminating dopamine signalling in the striatum (Budygin et al. 1999a; Budygin et al. 1999b).

Dopamine cannot be directly measured in human brain. However, animal findings are consistent with data obtained using positron emission tomography (PET). This showed greater D_1 receptor availability in the PFC of Val homozygotes, relative to Met allele carriers, which the authors interpreted as a measure for the lower PFC dopamine in Val homozygotes (Slifstein et al., 2008).

1.2.5.1. COMT and cognition

Given the importance of cortical dopamine in cognitive function, and COMT's regulation of PFC dopamine, much research has been conducted on whether the COMT polymorphism influences cognition.

Egan et al. initially found that executive control, measured by performance on the Wisconsin Card Sorting Test (WCST), was positively correlated with Met allele dose i.e. Met homozygotes performed best (Figure 5). However, subsequently, findings have been mixed and a large number of cognitive studies have not replicated the Met-allele advantage over Val homozygotes. Accordingly, meta-analytic findings are mixed (Barnett et al. 2007; Barnett et al. 2008).

More consistent is the finding that carriers of the less active Met variant have a higher PFC activity, for a given level of performance, in fMRI studies (Egan et al. 2001; Mier et al. 2010) and lower resting PFC functional connectivity (Tunbridge et al. 2013).

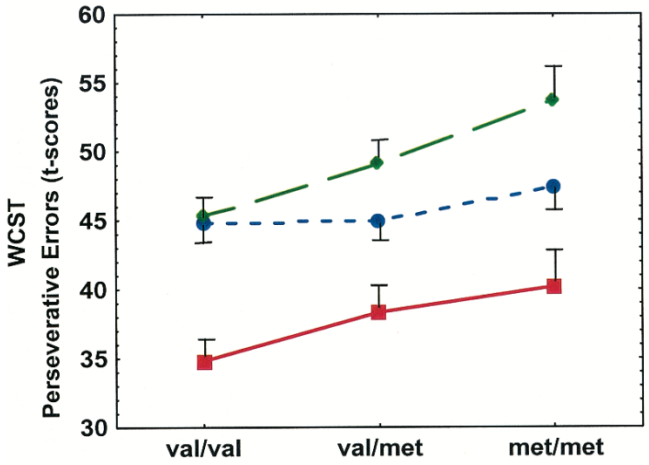


Figure 5: Association between WCST performance and COMT genotype. Met homozygotes performed better than Val carriers in patients with schizophrenia ■, in siblings ●, and in controls ◆ (lower scores indicate worse performance).
From (Egan et al. 2001).

Pharmacological modulation of COMT activity has shown more consistent effects on cognition. Inhibition of COMT by tolcapone in humans is associated with improved working memory (n-back task) and cortical processing in the dorsolateral PFC (Apud et al. 2007), and similar results are seen with amphetamine, an indirect monoaminergic agonist (Mattay et al. 2003) in line with the inverted-U relationship between cortical dopamine signalling and cognitive function (Figure 6).

The Inverted-U also predicts that the effects of COMT inhibition differ according to COMT genotype (Figure 6 right). Consistent with this, whilst Met-homozygotes outperformed Val-homozygotes subjects on the 2-back working memory task in those given placebo, this relationship was reversed in subjects given tolcapone and had opposite effects in the two genotype: it impaired performance in Met subjects but enhanced it in Val subjects (Farrell et al. 2012). The results are consistent with the Inverted-U model of dopamine function (Figure 6).

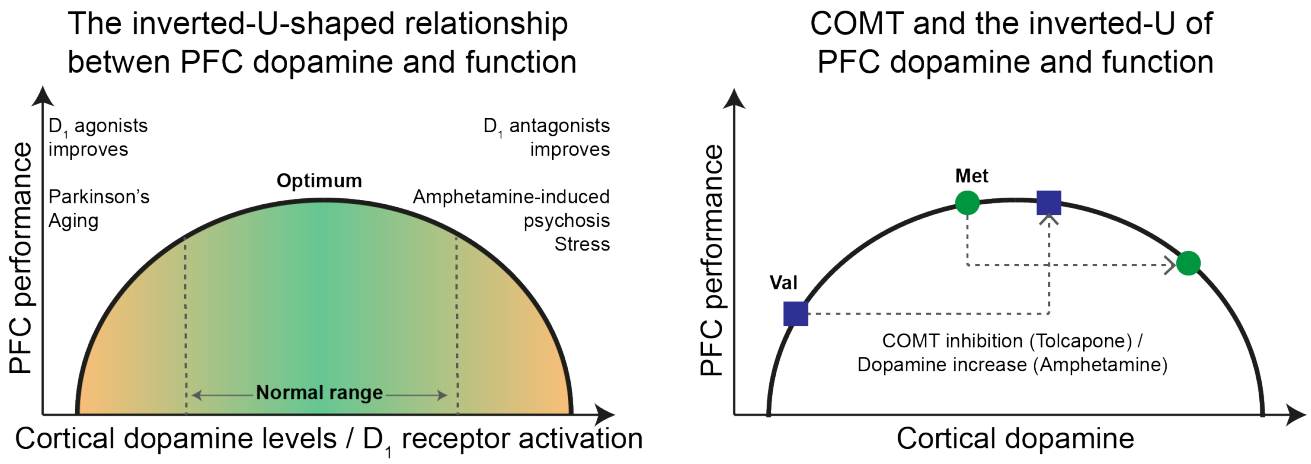


Figure 6: The inverted U-shaped relationship between cortical dopamine and PFC function, and its modulation by COMT. Left: PFC function is optimal at intermediate levels of cortical dopamine (green) and impaired in states (orange) of decreased (e.g. due to age or loss of dopaminergic neurons in Parkinson's disease) or elevated dopamine signalling (e.g. during stress or amphetamine-induced psychosis (adapted from(Goldman-Rakic et al. 2000)) Right: COMT Val-homozygotes have higher COMT enzyme activity and lower PFC compared to Met-homozygotes. They, therefore, occupy a position left of Met homozygotes on the Inverted-U. Pharmacologically induced increased in dopamine levels moves the subjects rightwards along the curve. Adapted from (Elizabeth M Tunbridge et al. 2006b).

Briefly, rodent studies also show cognitive improvements by COMT-inhibition. Tolcapone increases mPFC DA and improves extra-dimensional set-shifting (EDS i.e. a measure of cognitive flexibility) in (E. M. Tunbridge et al. 2004), spatial novelty preference (L. M. Laatikainen 2010), and delayed spatial win-shift test, single-trial passive avoidance test in rats (Khromova et al. 1997; L. M. Laatikainen 2010; Lapish et al. 2009). Studies in transgenic mice also support a relationship between COMT activity and cognition, as reviewed in Chapter 6.

Finally, whilst there is no evidence that dopamine in the ventral striatum is regulated by COMT, evidence suggests that COMT may nonetheless influence subcortical dopamine under certain conditions (E. M. Tunbridge et al. 2012), which may in turn influence reinforcement learning. These data are discussed further in Chapter 5.

1.3. Reinforcement learning and dopamine

How do we learn about the consequence of our actions to select those that lead to the most positive outcomes, whilst avoiding negative outcomes? Moreover, how are these processes influenced by genetic and environmental factors?

Psychological dual-process theories of decision making hold that there are two interacting valuation systems governing choice. In the first, goal-directed actions are driven by causal relationships between actions and outcomes, which are sensitive to changes in action–outcome contingencies and devaluation of the outcome. In contrast, habitual actions are driven by stimulus–response relationships and insensitive to changes in action–outcome contingencies e.g. outcome devaluation (Wood and Rünger 2015).

In cognitive neuroscience, the prevailing theory is that goal-directed actions and habits are two different, but interacting evaluation strategies grounded in reinforcement learning (RL): model-free and model-based RL (Daw et al. 2011).

1.3.1. Model-free RL and the striatum

Model-free learning uses reward prediction error (RPE) signals from midbrain dopamine neurons (Glimcher 2011; O'Doherty et al. 2017). The RPE signal is encoded by phasic activity of midbrain dopamine neurons projecting to the striatum (O'Doherty et al. 2004; Wolfram Schultz 2006) and represents the discrepancy between the reward received and that expected (Figure 7):

Reward Prediction Error = Reward Predicted – Reward Received

The RPE modulates the plasticity of the striatal targets so that actions leading to good outcomes are reinforced, whilst those leading to negative outcomes cease. Using RPEs, the model-free learning system progressively acquires cached estimates of the long-term values of circumstances and actions from retrospective experience. However, when learning from RPE signals alone the

value of an action is credited only by integrating across past reinforcements. Because the predictive value does not include the subject's motivation at the time of reinforcement, nor the identity of the reward (e.g., the colour or flavour, not the value in, e.g., metabolic or hedonic units), this way of learning is insensitive to changes in outcome values and inflexible to changes in the environment (O'Doherty et al. 2017).

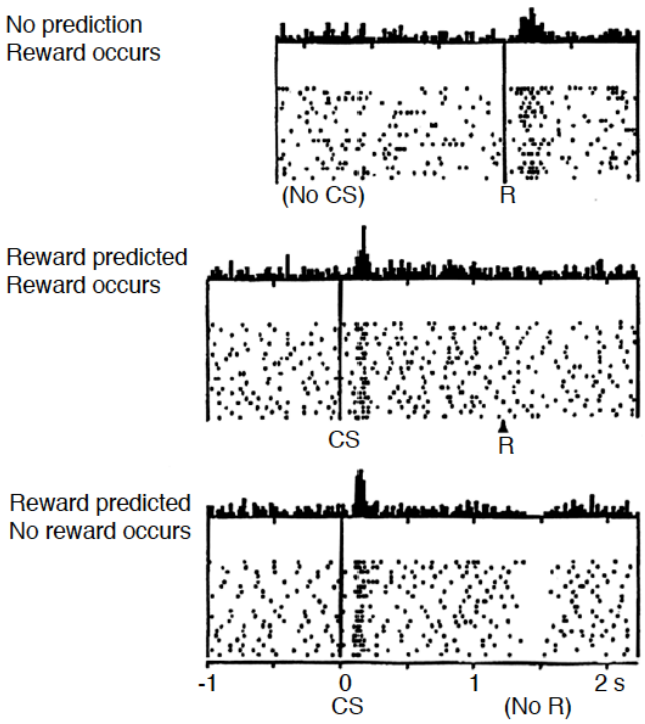


Figure 7: Changes in dopamine neuron firing during associative learning conditioning. An animal is trained to associate a distinct visual cue with reward delivery (R). Initially, a dopamine neuron burst fires only at the time of reward delivery (top). Once the animal has learned the association, burst firing is instead only elicited by the predictive cue (CS) and not by the reward (middle). The omission of a predicted reward leads to a dip in dopamine neuron firing (bottom). From (W. Schultz et al. 1997).

1.3.2. Model-based RL and the prefrontal cortex

In contrast to model-free RL, model-based learning is a flexible, but computationally demanding, process of internal stimulation and planning (Daw et al. 2005; Huys et al. 2012; Keramati et al. 2011),

In model-based RL, the animal learns a model of the world which predicts its future state given a chosen action and evaluates possible choices by simulating their consequences several steps into the future by generating an action-outcome tree or cognitive map of options and intermediate states. Although slow and computationally demanding, it uses evidence efficiently, as the implications of changes in the environment can be evaluated in the model without having to experience them directly (Daw et al. 2011).

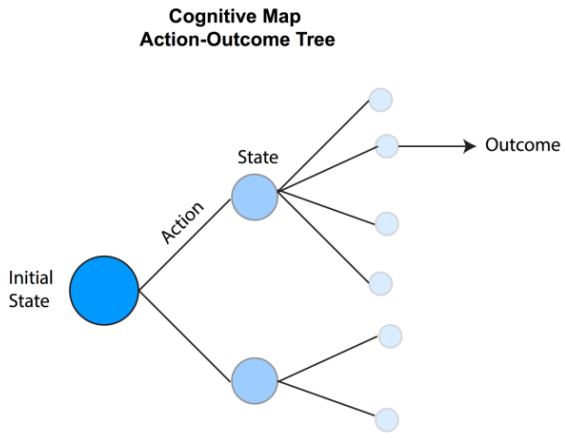


Figure 8: Cognitive map of an action-outcome tree. The map (the model) requires estimation of values of actions, states and outcome and a tree-search to identify the necessary sequence of actions and states to acquire the desired outcome.

Model-based RL relies heavily on the PFC (Daw et al. 2011; R. A. Otto et al. 2013c; R. A. Otto et al. 2013d) for value estimation, outcome identity, generating a cognitive map or action-outcome tree, and integrating all this information (O'Doherty et al. 2017).

Specifically, the dorsolateral PFC in humans and medial PFC in mice is associated with the manipulation of information during working memory (but not the maintenance of information in WM) (Wager and Smith 2003), and may be involved in the stimulation and planning of

forthcoming sequential actions based on information stored in short term memory (Pochon et al. 2001).The PFC, and cortical dopamine function, are therefore considered essential for model-based RL. Consistent with this proposal, disruption of the dorsolateral PFC by transcranial magnetic stimulation in humans impairs model-based learning (Smittenaar et al. 2013), while L-DOPA (dopamine precursor, Figure 2) enhanced it, and model-based learning-related dorsolateral PFC activation determined using fMRI (Deserno et al. 2015b). Moreover, Parkinson's patients off their dopaminergic medication show impaired model-based RL, which is restored on medication (Sharp et al. 2016), further implicating cortical dopamine in model-based learning. Also consistent with a role for cortical dopamine in model-based RL, a recent study found that the COMT Met allele dose dependently increased this form of learning (Doll et al. 2016).

1.3.3. Summary of model-free and model-based RL

Briefly, model-based decision making is goal-directed and flexible, but more computationally demanding, while model-free is faster and simpler requiring less integration from various brain regions, however making it more rigid and insensitive to changes in outcome value and contingency (Table 2). The neuronal resources utilised in the two systems are highlighted below.

	Model-based learning *Goal-directed action*	***Model-free learning*** *Habit action*
Cognitive demand	High	Low
Behavioural flexibility	High	Low
Sensitivity to change	Sensitive to changes in action-outcome contingency and outcome value	Stimulus-bound; insensitive to changes in action-outcome contingency or outcome value
Neural substrates	Dorsal cingulate cortex, Posterior parietal cortex, Hippocampus, ventromedial PFC, dorsolateral PFC, Orbitofrontal cortex, Basolateral Amygdala	Ventral striatum/ Posterior putamen

Table 2: Key characteristics of goal-directed and habitual instrumental actions. From (Lars Schwabe and Wolf 2011) and (O'Doherty et al. 2017).

The two learning systems are utilised in parallel, but arbitration between the systems during the final step of action selection is debated with some authors arguing a competition based on estimate accuracy or computational demand, or a top-down regulation (O'Doherty et al. 2017; A. R. Otto et al. 2013a; Wood and Rünger 2015).

Basal ganglia-thalamocortical circuits connect the basal ganglia with not only motor-related cortex areas but also frontal areas (G E Alexander et al. 1986). The circuits involved in the processes of motivation and cognition (Haber and Calzavara 2009), and the interaction between the two RL systems has been theorised to be mediated by these circuits (Dayan and Berridge 2014; Ito and Doya 2009).

Model-based and model-free RL systems have been implicated in multiple psychiatric phenotypes (Culbreth et al. 2016; Deserno et al. 2015a; Eppinger et al. 2013; Pizzagalli 2014; Radenbach et al. 2015; Sebold et al. 2014; Skatova et al. 2013; Voon et al. 2015b; Voon et al. 2015a), and even moral judgments (Ayars 2016).

Thus, a full understand of these processes has the potential to shed light on both healthy and aberrant brain function (Gillan et al. 2016).

1.3.4. The two-step task

This thesis employs a two-step probabilistic reward learning task to distinguish between model-based and model-free RL (Daw et al. 2011). The task, originally designed for humans with monetary reward has been adapted to rodents with water rewards (Thomas Akam et al. 2015) and (here) for use with sucrose rewards. A human version is described below, for illustrative purposes (Figure 9)(Sharp et al. 2016).

How the participant chooses on the first stage in response to events during the trial (i.e. outcome and transitions) can dissociate model-free and model-based choices (Figure 10).

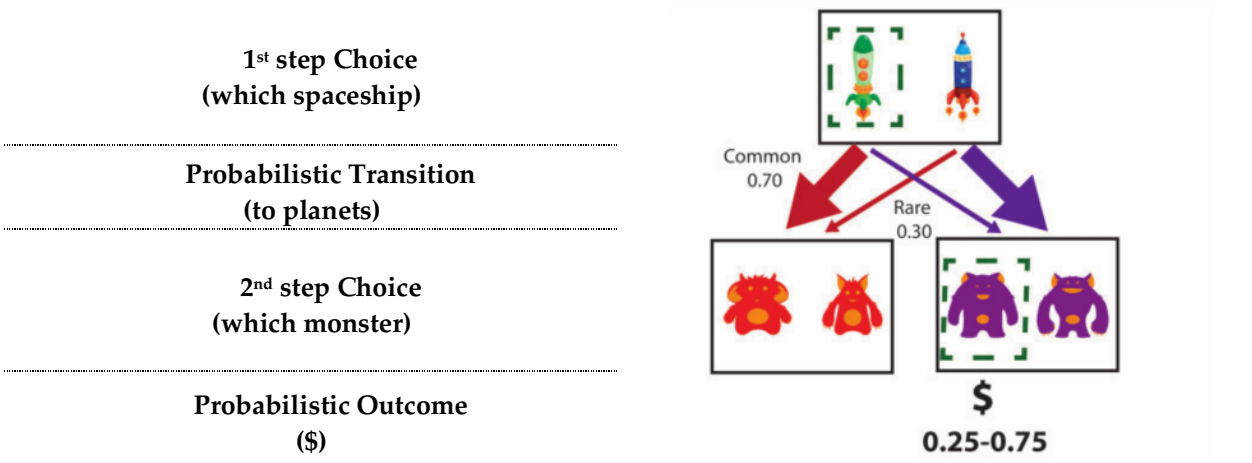


Figure 9: The human two-step task. The participant makes a 1st step choice (green or blue spaceship), and it transitioned commonly (70% of trials) and rarely (30%) to a specific state (planet), where the 2nd choice of (big or small monster) leads to a probabilistic reward ($). From (Sharp et al. 2016).

If a participant repeats the choices that are rewarded, disregarding the task structure, they are following Thorndike's "law of effect" which states that behaviour followed by pleasant consequences is likely to be repeated, while behaviour followed by unpleasant consequences ceases (Thorndike 1898): the evaluation strategy of a purely model-free agent. In contrast, a purely model-based agent with an internal representation of the task structure bases choices on that model and sticks with their conviction (i.e. the model), repeating the choice that most often leads to reward (i.e. the common transition to the specific state most often leading to reward).

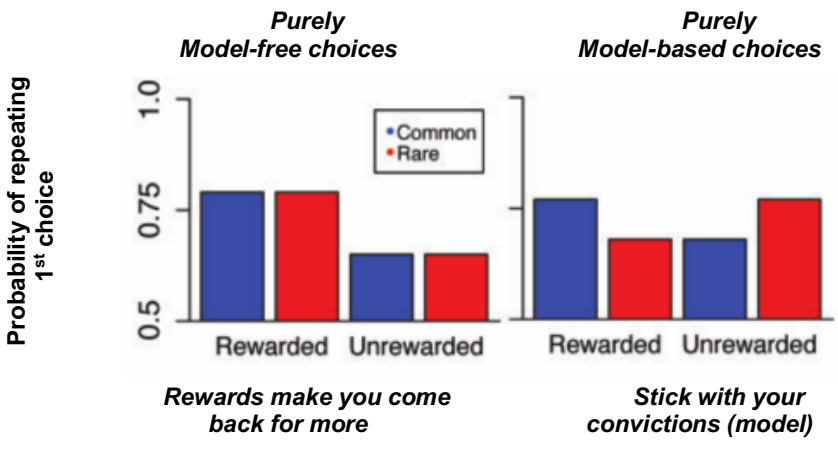


Figure 10: Dissociating model-free and model-based choices in the two-step task. Consider a trial where the 1st stage choice is followed by a rare transition (in Figure 9, to the planet less likely to follow that choice) and then reward. On the subsequent trials, repeating the same 1st stage choice (i.e. 'staying' with the same spaceship) indicates model-free learning (guided only by reward, and insensitivity to the probability). In contrast, making the alternate choice (i.e. switching to the other spaceship commonly associated with the planet that produced the reward) indicates model-based learning because the effect of reward is mediated by the probability structure. From (Sharp et al. 2016).

Using logistic regression, and a RL model, it is possible to dissociate the influence of the model-free and model-based evaluation systems on choice behaviour (Daw et al. 2011).

1.4. The acute stress response, COMT and RL

As outlined above, model-based decision making relies on cortical function, which in turn is modulated by COMT genotype. In addition, as outlined below, cortical function is modulated by acute stress (Amy F. T. Arnsten 2009b; A. F. T. Arnsten 2015). Therefore, I examined the impact of stress on rodent two-step performance, in both wild-type mice and those with altered COMT activity to study the interaction between the COMT polymorphism and acute stress on model-based RL.

1.4.1. Acute stress

In response to a stressful encounter, a physiological and neuronal response prepares the organism for an adaptive response. Hypothalamic neurons in the paraventricular nucleus (PVN)

release corticoid-releasing hormone (CRH), the hormone regulating the stress hormone cascade. CRH stimulates the corticotropic cells of the anterior pituitary to release adrenocorticotropic hormone (ACTH), which enters the systemic circulation and induces the adrenal cortex to release glucocorticoids. In humans, the primary glucocorticoid - or 'stress hormone' - is cortisol, whilst it is corticosterone (CORT) in rodents. As soon as glucocorticoids and other steroid hormones are made in the adrenal glands, they are released. Corticosteroids affect brain and body functions through both rapid (non-genomic) and delayed (genomic) mechanisms.

Sensory information relating to physiological stressors (i.e. those that involve an immediate potential threat, e.g. a sudden loud noise) is communicated to the PVN by first- or second-order neurons, generating a direct activation of the hypothalamic–pituitary–adrenal (HPA) axis (Herman et al. 2003).

Psychological stressors (such as the worries of embarrassment or inadequacy experienced e.g. during job interviews, as simulated by the widely-used Trier Social Stress Test (TSST)(Kirschbaum et al. 1993) require interpretation and activate the HPA axis indirectly, by PVN stimulation through the limbic system (Herman et al. 2003).

1.4.1. Acute stress and model-based/model-free RL

Acute stress impairs model-based decision making in humans (R. A. Otto et al. 2013d; Radenbach et al. 2015). However, high working memory capacity protects against the stress-induced model-based impairment (R. A. Otto et al. 2013d). Model-based RL is also impaired in patients with obsessive compulsive disorders and schizophrenia, alcohol and amphetamine addiction, and in old age (Culbreth et al. 2016; Sebold et al. 2014; Voon et al. 2015a).

More insight into disturbances or protective effects on model-based RL might, therefore, give insight into the neurobiology underlying for different mental states, as well as a shared framework for their interpretation.

1.4.2. COMT and cognitive function under acute stress

The PFC has abundant glucocorticoid receptors (see 1.4 below) and is sensitive to both acute and chronic stress, which can lead to prefrontal dopaminergic dysfunction and impairment of cognitive function (A. F. Arnsten 2009a; A. F. T. Arnsten 2015). Acute stress elevates cortical dopamine (Amy F. T. Arnsten 2009b) and (similar to the inverted-U of PFC dopamine and cognition), an earlier inverted-U of stress and PFC function has also been proposed (Sapolsky 2015; Yerkes and Dodson 1908).

As outlined above, the promoter region of the COMT gene has hormone binding sites that regulate COMT expression (Al-Hendy and Salama 2006; Xie et al. 1999), which may contribute to the COMT dependent effects of cortisol responsiveness and stress (Armbruster et al. 2011; Thompson et al. 2012). Several studies have demonstrated that the COMT polymorphism impacts the cognitive, neural and physiological responses to acute stress (Buckert et al. 2012; Qin et al. 2009; Radua et al. 2014), and suggested that the inverted-U relation between stress and PFC dopaminergic functioning is COMT-dependent (Hernaus et al. 2013).

Acute stress negatively impacts working memory in Met subjects, whilst improving it in Val subjects (by Trier Social Stress Test: (Buckert et al. 2012); by traumatic film: (Qin et al. 2009)). This interaction was also observed for PFC activation (stress increased working memory-related activation in the PFC in Val carriers but decreased it in Met homozygotes; Figure 11 (Qin et al. 2009). A PET study also found less PFC dopamine release (by ([^{18}F]fallypride displacement) in Met carriers compared to Val homozygotes in response to psychosocial stress (Montreal Imaging Stress Task); interestingly Met carriers also reported greater subjective feelings of stress than Val homozygotes/carriers (Hernaus et al. 2013). COMT mouse models also show altered stress responses (see in Chapter 6).

However, the possible interaction between COMT genotype and acute stress on model-based decision making has not been studied, and is the subject of this thesis.

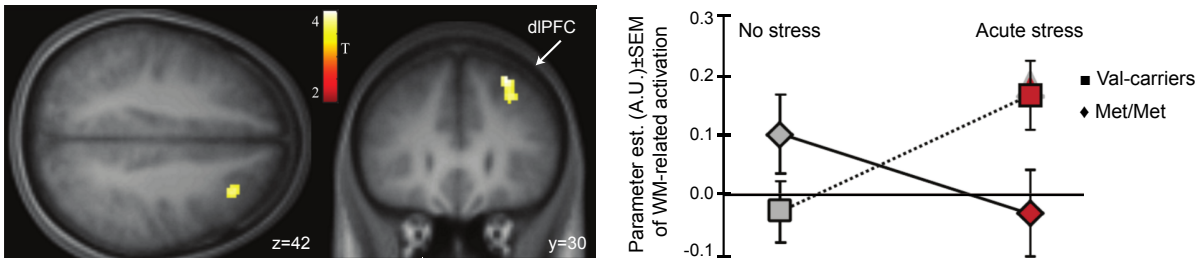


Figure 11: Acute stress and COMT-genotype interact to affect WM-related (*n*-back task) activity in the dorsolateral PFC. Left) Transverse and coronal views of activation in the right dorsolateral PFC (dlPFC) showing a stress-by-COMT genotype interaction (statistical parametric maps threshold at $P<0.001$). Right) Bar graphs of parameter estimates±SEM of the WM-related activation under control and stress (aversive movie clips) in Met homozygotes and Val-carriers. Adapted from (Qin et al. 2009).

1.5. Hypotheses

The aim of this MSc thesis is to investigate the effect of COMT genotype, and acute stress, on model-based and model-free RL.

As detailed above, the rationale for this research is based on four main points:

1. COMT genotype modulates cortical dopamine and function (section 1.2.5)
2. Model-based decision making relies on cortical function (section 1.3.2)
3. Cortical function and model-based learning are impaired under acute stress (1.4.1 & 1.4.2)

The information presented above suggests the COMT Val^{158}Met polymorphism modulates cortical dopamine and function, and could, therefore, modulate RL decision-making under acute stress.

As this study is largely based on computational analysis of choice behaviour, I will start by explaining the behavioural task, the analyses employed, and how I expanded or modified them in Chapter 2, the General Methods.

The task had to be adapted from water to sucrose rewards, as detailed in Chapter 3, in which I validate the task and compare the two reinforcement protocols.

Chapter 4 studies the effect of acute stress on model-based and model-free RL in wild-type mice.

Chapter 5 studies the separate and interactive effects of COMT genotype and acute stress on model-based and model-free RL in a mouse model of the human COMT Val^{158}Met polymorphism.

Finally, Chapter 6 provides a general discussion of the results of this thesis and suggests further directions for research.

2. General Methods

All experiments and procedures were approved by the Home Office and performed according to the Animals (Scientific Procedures) Act 1986 and associated guidelines, under the auspices of the ethical review board at the University of Oxford. Data from experiments and procedures carried out abroad by colleagues adhered to EU directive 2010/63/EU on the use of animals for scientific purposes.

2.1. Mice

A total of 40 mice were used for these experiments (see below for precise details). Mice were housed in cages of 2-4 littermates with paper-based nesting material, boxes, and tubes. On arrival at the animal facility, the mice were acclimatised for one week followed by approximately a week of food restriction before training. The animal facility was maintained at 19-23°C.

All mice were maintained at ~85% of their free-feeding weight on 2016 Teklad Global 16% Protein Diet (Evigo). Free-feeding weight was measured at body score 3 and during food restriction their body score was 2-3. On non-testing days, the mice were weighed and fed at approximately the same time as the previous testing day. The mice were tested 5-7 days/week during the light phase of a 12-hour light/dark cycle.

2.1.1. Pilot cohort (SUCROSE 1)

This initial cohort, used to pilot and optimise the task, consisted of male C57BL/6JOlaHsd (Evigo) mice (n=8) held in two cages. One mouse was culled day 13 due to fighting with cage mates, leaving n=7 to complete the training. Testing began at ~11 weeks of age. Testing took place between 9:30 am - 1:30 pm during the first 35 days and 10:45 am – 4:30 pm during the later drug screening.

2.1.2. Second cohort (SUCROSE 2)

The impact of injection stress on task performance was examined in the second cohort of male C57BL/6JOlaHsd (Evigo) mice. Sixteen mice, held in four cages, were initially trained for 6 days. The most active mice in the task at this point (i.e. those with the highest cumulative activity) were selected for the main task training (day 7, n=8). The unselected mice remained in the home cages so as not to disturb the established social structure. Testing started at ~12 weeks of age. Sessions took place between 9:30-12:15, 5-6 days a week during training and 7 days a week during the injection stress study.

The cohort was used for a larger pharmacological study of the effect of the COMT inhibitor Tolcapone and the DAT inhibitor GBR-12909 (or Vanoxerine), which is not part of this thesis and not presented further. The study started on day 39. The mice received an intraperitoneal (i.p.) injection with Tolcapone (30 mg/kg in vehicle) or vehicle alone 90 min before testing, or and GBR-12909 (6 mg/kg in vehicle) or vehicle alone 15 min before testing. Vehicle was 20% cyclodextrin in saline. Injection volumes were 10 ml/kg.

Injections were performed using a 26-gauge needle in a room separate from the testing room, volumes were 10ml/kg and the injection side (lower left vs. right abdominal quadrant) was alternated each injection day, to avoid tissue scarring and irritation. Injections were counterbalanced in a Latin square design by day and injections in 6 day cycles (page 140). A day without injection was interleaved between each day of injection.

2.1.3. COMT cohort

We used a mouse model of altered COMT activity (Barkus et al. 2016). Briefly, in the COMT-Met mice the native Val158Met-equivalent COMT amino acid (Leu148) was replaced with a methionine by a PCR-based strategy by J. Chen and D. Weinberger at the National Institute of Mental Health and the Lieber Insitute for Brain Development, United States (Appendix). This

resulted in homozygotes (COMT-Met mice, which show low COMT activity (Barkus et al. 2016)) and can be compared with their (high COMT activity) wild-type (COMT-WT) littermates. Genotype was confirmed by polymerase chain reaction (PCR, Appendix). Male, homozygous COMT mice (n=8 WTs; n=8 COMT-Met) were divided into two groups of eight (n=4 COMT-Met and n=4 WTs per group, referred to as groups COMT.1 and COMT.2). These two groups were run in parallel under the same conditions; COMT.1 were always tested between 12:15 pm – 3:30 pm and COMT.2 between 2:45 pm – 6 pm. During the stress study the test starting time was more stringently held to 12:15 pm – 2:30 pm and COMT.2 at 3:30 pm – 5:15 pm. The cohort was on average 20.3 weeks old at the start of training, which was balanced between genotype and group. (Table 3).

Group	**COMT cohort**	**COMT.1**	**COMT.2**	**COMT-WT**	**COMT-Met**
Age (mean±SD)	20.3±1.2	19.8±1.5	20.8±0.4	20.2±1.0	20.4±1.3

Table 3: Mean age of the COMT cohort divided by group and genotype.

2.1.4. Mice on water deprivation (WATER 1 and WATER 2)

To assess whether the task could be adapted for use with sucrose reward from the original version (which used water reward), we compared our data obtained in food-restricted mice with sucrose reinforcement with those previously obtained in water-deprived mice with water as the reward (two cohorts of C57BL/6, n=6, 10.0±0.8 weeks old with 39 testing days and n=6, 11.0±0.6 weeks old with 45 testing days). These data were acquired by Thomas Akam and colleagues at Fundação Champalimaud, Portugal, between June 2014 and June 2015 (Akam T. 2016) and in accordance with Portuguese animal research legislation and EU directive 2010/63/EU. Testing usually started at 3-5 pm and session duration was 2 hours. During water deprivation, the animals only had access to water during the task and an hour *ad libitum* during days without testing.

2.1.5. Power and sample size

We set a significance level of α=0.05. A power of 80% was aimed for; however, the rodent two-

step task is a newly developed task that has only been used to study within-session changes of behaviour during optogenetic inhibition of the anterior cingulate cortex in small groups of animals (Akam T. 2016) and so formal *a priori* power calculations were not possible.

2.2. The two-step reinforcement learning task

2.2.1. Testing boxes

Testing was conducted in custom-made Plexiglas boxes (14 x 14 x 14 cm, Figure 12). Mice make a series of nose pokes to receive probabilistic rewards of sucrose solution. The back wall had four nose-poke ports. The lateral walls and the wall with the nose pokes were black, while the roof and back wall were transparent. Each poke was surrounded by four LEDs that lit up as cues during the first and second steps (Figure 12, and on page 141). Each box had a strip of three LEDs covered by black paper to provide light in the boxes during testing.

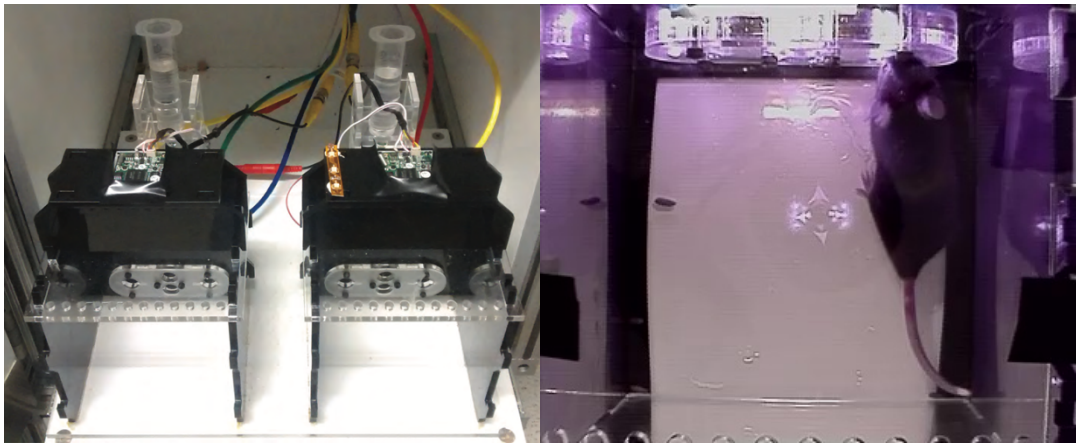


Figure 12: (Left) The custom-made testing boxes. (Right) A mouse receiving reward in lateral poke hole. The left image is from Champalimaud where the cameras are mounted directly on top and thus differ slightly from those employed in Oxford. Due to security policy, no image is available from Oxford.

Activity was measured by infrared beam breaks in the nose-poke holes and recorded by the pyboards. Behaviour was observed and recorded by 700TVL Bird Box Cameras (SpyCameraCCTV) and saved on an SRD-876D digital recording device (Samsung). The boxes were held in sound-attenuating cubicles, equipped with fans (Med Associates). Delivery of rewards, 20% w/v sucrose solution (Sigma-Aldrich) was controlled by solenoids (Jaltek Systems),

which, together with the infrared beam detectors and LED lights, were connected to pyboards (Jaltek Systems) running custom Python 3.5 code (T. Akam 2016a). The solenoids, HDI valves (Lee Products), were calibrated individually. After and between animals during testing, the boxes were cleaned with Anistel High Level Surface Disinfectant (Tristel). After testing the solenoids and tubes were flushed with boiled distilled water. Approximately every fortnight solenoids and tubes were flushed with 0.6 % v/v Milton Sterilising Fluid (Laboratoire Rivadis) in distilled water for 15 min followed by distilled water.

2.2.2. Task structure

We employed an appetitive reinforcement learning task termed the 'two-step' task which is designed to distinguish between model-based and model-free reinforcement learning. The task is adapted from a behavioural paradigm originally developed for humans (Daw et al. 2011).

Each trial of the task consists of a choice between two 'first-step' actions, which leads probabilistically to one of two 'second-step' states, where a further action gives a possibility of reward (Figure 4b). Both the reward probabilities in each second-step state, and the transition probabilities mapping the first-step actions to the second-step states, change over time. To obtain rewards at a high rate the subject must, therefore, choose that first-step action which currently leads with high probability to the second-step state which currently has high reward probability.

The task is implemented in operant boxes with four nose-poke holes each with its own stimulus LED light; a high and low poke in the centre, flanked on either side by a left and right poke. Each trial starts with the central (high and low) pokes lighting up. The choice of high or low poke causes the central poke stimulus lights to go off and either the left or right poke to light up. The subject then pokes the illuminated side poke for a probabilistic reward of sucrose solution (4 ul of 20%), followed by a 1-second inter-trial-interval (ITI). If the mouse erroneously poked into a non-active (not lit) hole, the activity was registered but did not lead to a change in the task or a reset the

trial. The trials only ended when the mouse eventually poked the active (lit) side hole. The task is self-paced so that the subjects can perform as many trials as they want within a session and are not directly penalised for taking time out.

The rodent task is adapted from a behavioural paradigm originally developed for humans (Daw et al. 2011) and simplified with *a*) enhanced contrast between good and bad options and *b*) no choice at the second step, Figure 16 (Akam, Costa, and Dayan 2015).

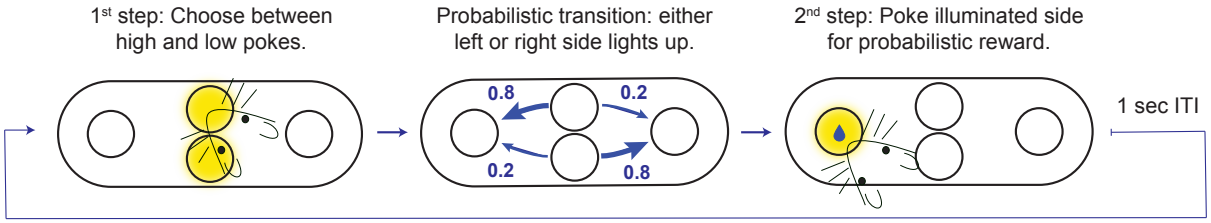


Figure 13: The trial structure of the two-step task for rodents. In the 1st step the central holes light up, and the subject chooses high/low, which then transitions to left or right, where the subject pokes to receive a reward. The inter-trial interval (ITI) was 1 sec.

Both the transition probabilities linking the 1st step choice (high or low) to the 2nd step state, (left or right side lit up) and the reward probabilities in each 2nd step state, changed in blocks. Each block was defined by the state of both transition and reward probabilities and, at a block transition, either the transition or the reward probabilities changed. The two possible states of the transition probabilities are A: *high-left* / *low-right* where high→left and low→right transitions were common occurring with 0.8 probability, and B: *high-right* / *low-left* with the reverse transition probabilities (Figure 14). The reward probabilities could also be in one of two states *left-good* with 0.8/0.2 reward probability on the left/right side, and *right-good* with 0.2/0.8 reward probabilities. Therefore, e.g. in a block with *high-left* / *low-right* transition probabilities and *left good* reward probabilities, the optimal choice is to make high initial (1st step) pokes to achieve the greatest chance of transitioning to the high reward probability side. With the transition and reward probabilities defined above, choosing the 'correct' option at the first step yielded reward on 68% of trials, whereas the other 'incorrect' choice yielded reward on 32% of trials.

While this task structure was used for the majority of the studies presented in the thesis, the original paradigm used for the water-reinforced mice in Lisbon and for the sucrose-reinforced Pilot cohort during the first 29 days of testing (Chapter 4) also included a neutral reward block, in which the reward likelihood was 0.4 for both left and right and with a fixed block length of 40 trials. The aim of the neutral blocks was to make it more likely the subjects track the reward probabilities as a continuous variable rather than inferring they are in one of two different states. They were removed from the Pilot cohort (with no apparent effect on behaviour) see page 57 and were therefore not used in subsequent studies.

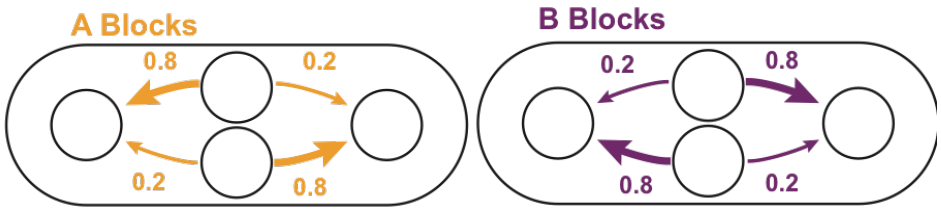


Figure 14: Transition blocks A and B determine common (80%) and rare (20%) transition from the central poke to the side which is subsequently illuminated.

The subject can erroneously enter an inactive poke hole e.g. during the 1 sec ITI or after the 1st choice in an unlit poke. Such pokes do not change the state of the task, and for the analyses reported here, we only analyse nose pokes to active options.

Block transitions were triggered based on a behavioural criterion, occurring 20 trials after an exponential moving average (τ = 8 trials) of 1st step choices reached over 75% correct. At block transitions, it was determined pseudo-randomly whether the transition or reward probabilities reversed such that, for every three block transitions, one was a transition probability reversal and two were reward probability reversals (Figure 15). Each session started with the same state of the transition and reward probabilities as the previous session finished on (exepct during an error, see page 57).

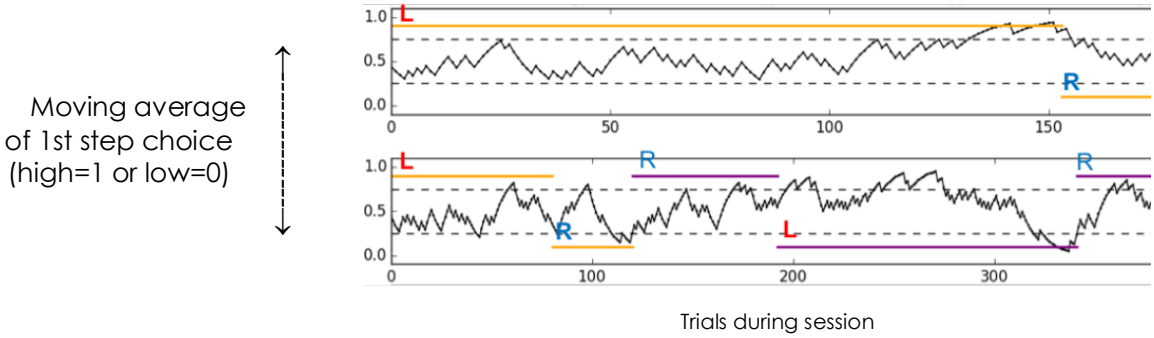


Figure 15: Plots of the initial half of sessions and first choices of two subjects. The black solid line shows the exponential moving average of the 1st step choice (high/low). Orange and purple bars indicate transition blocks A and B, respectively. R and L indicate reward blocks where either right (R) or left (L) is best. Dashed lines at 0.75 and 0.25 indicate the threshold for the transition. When the moving average goes above or below the threshold, the block changes stochastically after an additional 20 trials. (Top) The mouse is in block A with Left being the most rewarded option during the first half of the session. Therefore, the is the correct choice. The mouse alternates between choosing the top and bottom for the first 100 trials. It then builds a preference for the top and, at ~trial 130, reaches the threshold and persistently choices the top. However, at trial ~150 the reward block is changed, and the bottom choice became optimal. (Bottom) Example of a very fast learner who rapidly completed several block transitions during a session, indicating fast reversal learning. The reversal learning speed can be analysed and stratified by transition type (i.e. transition or reward block) in the reversal learning analysis.

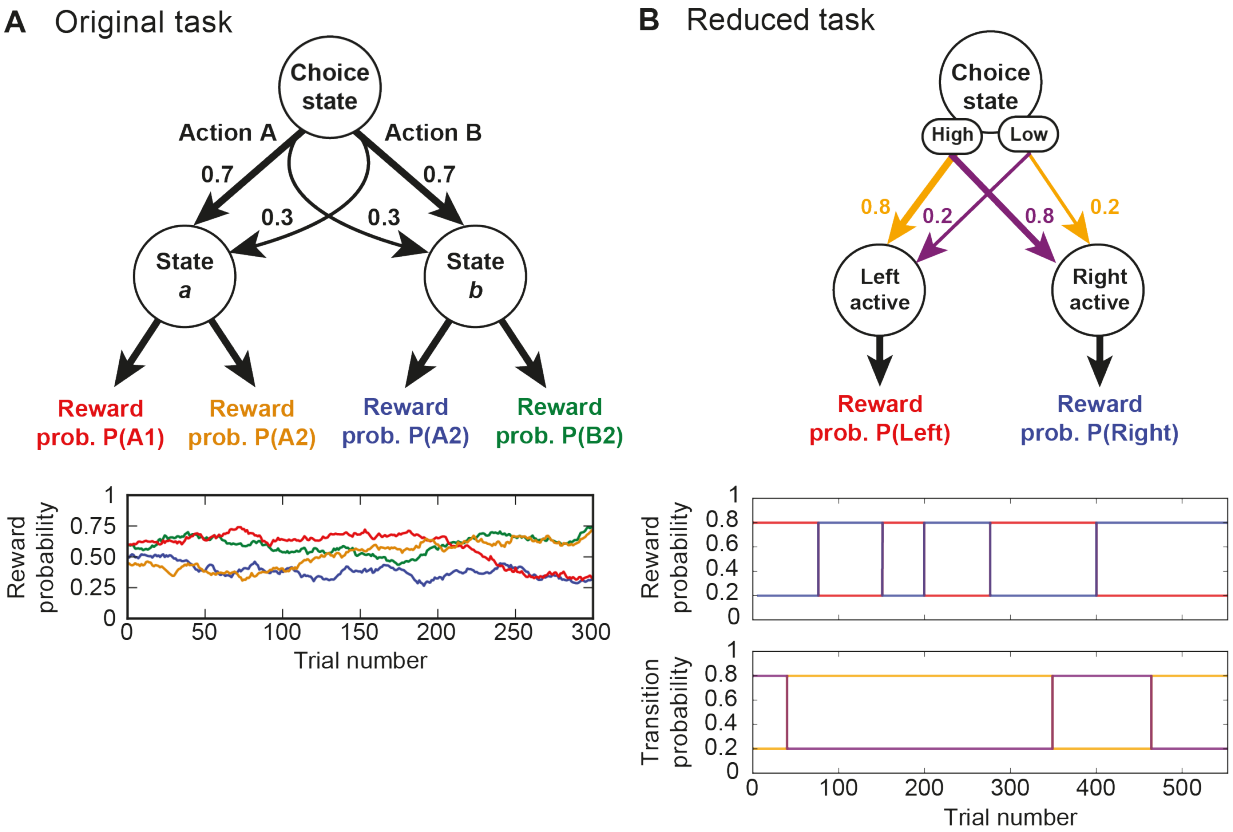


Figure 16: Comparison of (A) the original and (B) the reduced two-step task. Note the enhanced contrast (0.7/0.3→0.8/0.2) between the transition from initial choice state to the 2nd state and the removal of a secondary choice.

2.2.3. Task training

At the start of training, in order to create a high contrast between good and bad options, the transition probabilities were set to 0.9/0.1 (common/rare) and the reward probabilities to 0.9/0.1 (good/bad side). Both probabilities were changed later in training to 0.8/0.2: the transition probabilities were changed when the subjects performed on average >250 trials/session and >3 reversals/session; the reward probabilities were changed when the subjects again completed >3 reversals/session.

In the first session for each cohort, the transition and reward block was randomly assigned and the first 40 trials were all rewarded. During training, the reward size was gradually reduced from 8 µl to 4 µl sucrose solution. The session length was initially 1.5 hours, increased to 2.0 hours and, finally, 2.5 hours.

Training for the cohorts was considered finished when mice were running the full task, i.e. high transition contrast and high reward contrast, and the group average reached the following criteria:

1. A 3-day average of >450 trials/session
2. A 3-day average of >4 reversals/session
3. A 3-day average combined reversal learning speed (τ) of <30 trials

Criteria 1 and 2 ensure that mice are completing a sufficient number of trials and blocks per session to provide enough data for analysis of the effect of manipulations of interest, whilst criterion 3 ensures they are performing the task to a sufficient level of performance. More detail on training for each specific cohort is described in chapters 3 - 5.

2.3. Analysis

To measure the behavioural effects of COMT genotype and stress, the following analyses were used: reaction times, trial rates, choice probability trajectories following reversals, a logistic regression of choice behaviour, and a reinforcement learning (RL) model of choice behaviour.

Permutation tests were used to determine statistical significance for differences between groups in reaction times, trial rate, reversal learning, and the parameters of the logistic regression and RL model (see page 52). The analysis was largely conducted using custom code in the Anaconda package Python 3.5 (T. Akam 2016b) with NumPy (Walt et al. 2011) and Matplotlib (Hunter 2007) (except where stated otherwise).

2.3.1. Reaction times and trial rates

Reaction time from the 1st step choice (high or low central poke) to entering the active 2nd step poke (left or right) was analysed. The median reaction time per subject was stratified based on whether the transition was common or rare.

We used the trial rate and cumulative trials during the session as a measure of motivation.

2.3.2. Reversal analysis

The reversal analysis evaluated choice probability trajectories around block transitions, i.e. around reversals in the reward or transition probabilities. Two measures of performance were extracted from these trajectories:

1) The average probability of choosing the correct option in the 15 trials preceding a reversal, termed the ***'block end choice probability'*** (Figure 17, arrow). Higher values indicate that subjects are choosing the correct option more at the end of blocks and hence represent better performance.

2) The ***time constant τ of an exponential fitted to choice probability*** trajectory over the 40 trials

following a reversal (Figure 17, blue line). This provides a measure of how rapidly subjects learn following changes in the task structure, with lower τ indicating better performance. We analysed the τ separately for reversals in the transition and reward probabilities, and for both reversals combined.

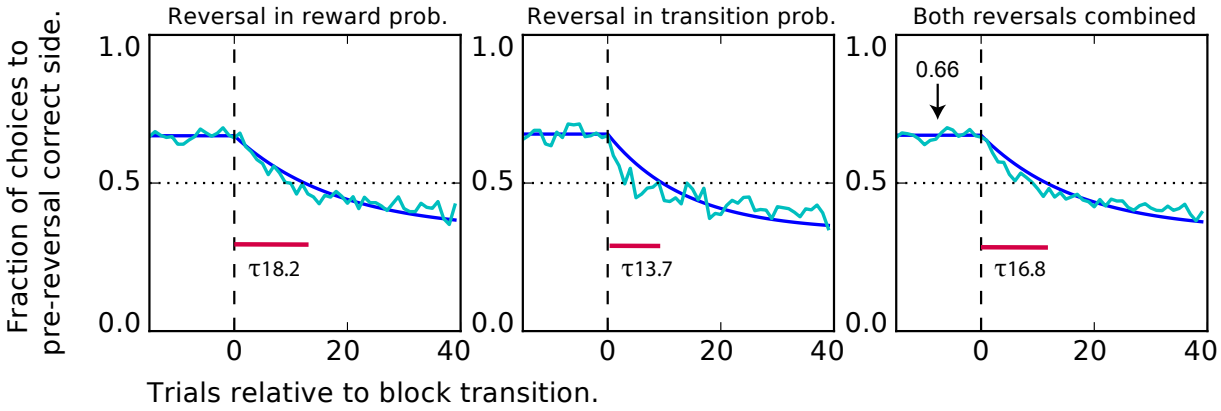


Figure 17: Example of reversal plot divided by the two types of block changes: transition A/B or reward Left/Right. Fit (blue) of choice behaviour (cyan) of the 15 trials before and 40 trials after a block change. Reversal τ here is shown as red lines and average block end choice probability (as indicated with an arrow).

2.3.3. Multilevel modelling

We used both a logistic regression model and a reinforcement learning model (described below) of subjects' choices to study their behaviours. Both approaches used multi-level modelling (also termed hierarchical or nested modelling) in which parameters were assumed to be Gaussian-distributed across the population of sessions, and the mean and variance at the population level was fitted using maximum likelihood methods (Figure 18).

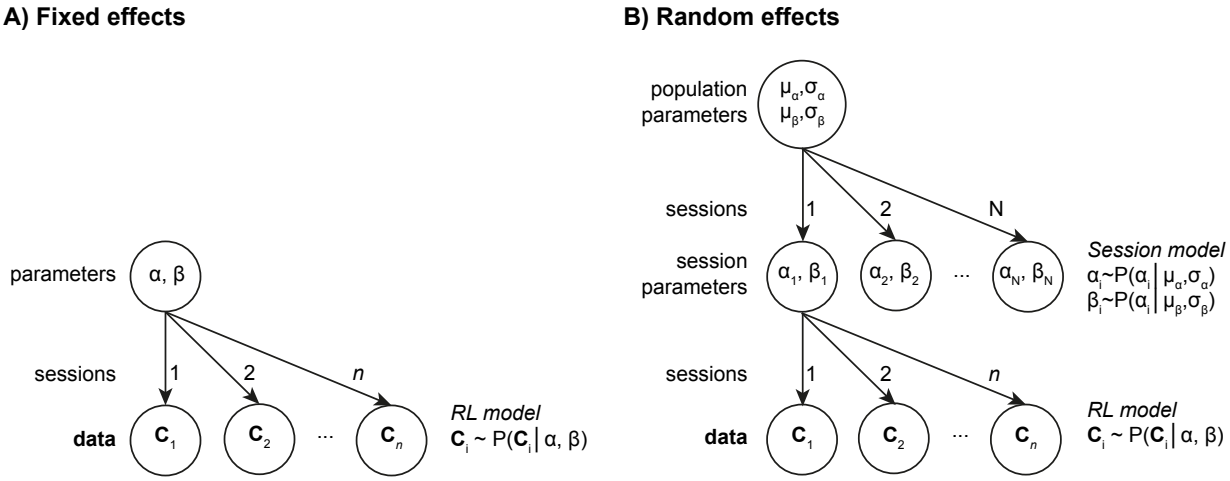


Figure 18: Models of population data: A) Assuming fixed effects, the model parameters are shared between subjects and ignores variability of subjects and sessions. Such an analysis treats data from all subjects as if it were a single subject. B) Using random effects, each session's parameters are drawn from a common population distribution. This creates a two-layer hierarchy of data (population and session) that acknowledges that the data (choice c_i) is acquired from individual sessions. Adapted from Delgado et al. 2011.

2.3.4. Logistic regression

The logistic regression model asked how events on one trial, specifically the outcome (rewarded or not), transition (common or rare), and their interaction, affected the probability of the subject repeating their previous choice. The motivation for this analysis is that different behavioural strategies produce different patterns of loading on relevant regression predictors (Thomas Akam et al. 2015; Daw et al. 2011).

The regression model predicts the probability of choosing the high-poke, as a function of a set of predictors. In defining the values taken by the predictors, we use the following variables: *C* for stay, *I* for outcome, *T* for transition and *R* for reward. They were coded as:

C = +1 if the previous choice was a high poke, -1 if a low poke.

O = +1 if the previous trial was rewarded, -1 if not rewarded.

T = +1 if the previous trial had a common transition, -1 if a rare transition.

R = +1 if the previous choice was the correct option (higher reward probability),

-1 if the previous choice was the incorrect choice (lower reward probability).

The model used the following predictors:

- **Bias** (high/low): +1 for all trials; tendency to choose high poke
- **Side bias** (clockwise/counter-clockwise): +0.5 if the previous trial ended on the left, -0.5 if it ended on the right; tendency to choose high returning from left and low from right
- **Correct**: 0.5 *C R*; tendency to choose the correct option at 1st step (the high/low option that commonly leads to the state with the highest reward probability)
- **Choice** (stay): 0.5 *C;* tendency to repeat 1st step choices
- **Outcome** (rewarded): 0.5 *C O*; tendency to repeat 1st step choices following *reward*
- **Transition** (common): 0.5 *C T*; tendency to repeat step 1 choices following *common* transitions
- **Transition x Outcome Interaction**: 0.5 *C T O*; tendency to repeat choices following rewarded common transition trials *and* non-rewarded rare transitions trials

The *outcome, transition* and *transition-outcome interaction* predictors are informative about behaviour strategy (see below). The other predictors are included to capture biases and perseveration, and are required to avoid spurious loading on trial event predictors (Akam T. 2016). The first 20 trials after a transition probability reversal were excluded as at that point it is ambiguous which transition was common and rare, which resulted in ~9% of trials being excluded from the analyses (Akam T. 2016).

To evaluate how different behavioural strategies appear to the regression analysis, data were simulated from a model-free and a model-based RL agent and analysed with the regression analysis (Figure 19 right). Data simulated from a model-free agent showed positive loading on the outcome predictor, i.e. it repeated rewarded choices, but had near zero loading on the transition and transition-outcome interaction predictors(i.e. it was insensitive to the experienced state

transition) (Akam T. 2016) (Figure 19). Data simulated from a model-based agent also showed positive loading on the outcome predictor but, unlike the model-free agent, showed positive loading on the transition predictor, indicating that it was more likely to repeat choices following common state transitions.

Experimental data from mice performing the task, showed loading on both outcome and transition predictors indicating that they use model-based control as part of their behavioural strategy (Figure 19).

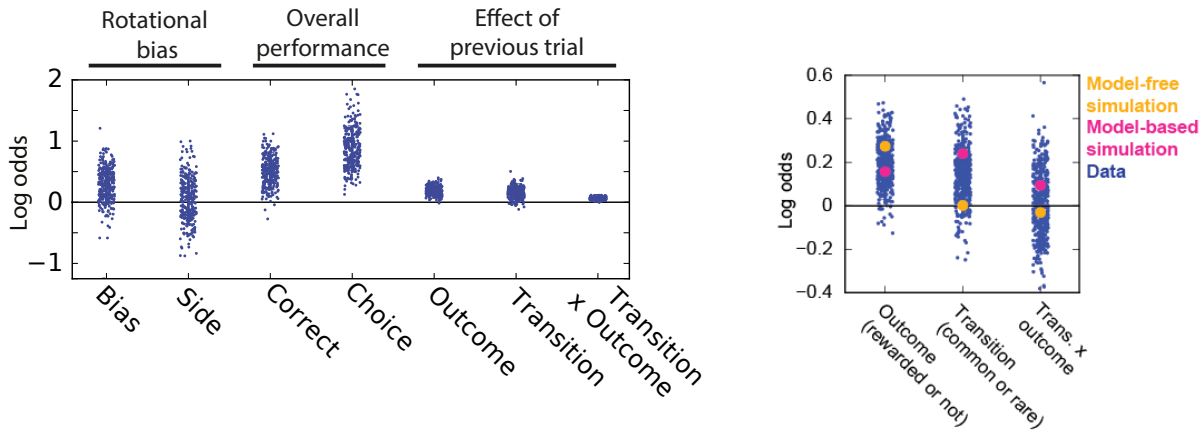


Figure 19: Example of logistic regression of choice behaviour. The dots represent maximum a posteriori session fits and the bars the standard deviation of the population level distributions. Due to the vertical position of the poke holes, the subjects consistently had a bias toward high. Boxes with higher holes showed less bias (personal communication. May 2016, T. Akam). Right: Subjects use a mixture of both model-free and model-based strategies (Water dep. data, blue). A simulated purely model-free agent (orange) shows loading on the outcome predictor, indicating a tendency to repeat rewarded choices, but does not show loading on the transition or interaction predictors. A model-based agent (magenta) shows loading on both the outcome and transition predictors, which indicates that both receiving reward and experiencing a common transition is reinforcing. The data indicates that subjects tend on average to use a mixture of both model-free and model-based strategies. Images from (Akam T. 2016).

2.3.5. Computational Reinforcement Learning Model

A second approach used to study the subjects' behavioural strategy was to fit reinforcement learning models (RL models) to their behaviour. An RL model aims to describe the learning and decision processes underlying behaviour as a set of equations, which represents how decision variables, such as action values, are updated in light of new experience, and how they guide future

choice. The models can be fitted to the subject's actual behaviour by adjusting their parameters to maximise the likelihood that the model would generate the observed behaviour. Different models can be compared in terms of how well they describe the data, by evaluating how likely they would be to produce the observed behaviour.

The RL modelling of the two-step task behaviour built on model-comparison work done by Akam (2016) following Daw (2011). Both had established that behaviour on the task was well-described by models that combine a model-free RL component and a model-based RL component, which each learned values for the actions. The action values learned by each system were then combined to give net action values which determine choice behaviour.

Figure 20 shows a schematic of the processes which update action values in the model-free and model-based components of the agent.

The model-free component learns action-values directly using RPEs. RPEs signal the difference between the expected and the actual (experienced) value of the outcome of an action. After each action is taken the RPE is used to update the agent's estimate of that action's expected value towards the experienced value of the outcome. In the two-step task model, RPEs occur when the subject makes the 1st step choice and reaches one or other 2nd step state (each of which has an estimated value associated with it), and again following the 2nd step action which obtains (or does not obtain) a reward.

In contrast, the model-based component learns a model of which state follows which action, updating these estimates using state prediction errors, i.e. the difference between the state prediction and experienced next state. The state predictions are combined with estimates of the state values (learned through RPEs) to give estimates of the action values at the 1st step.

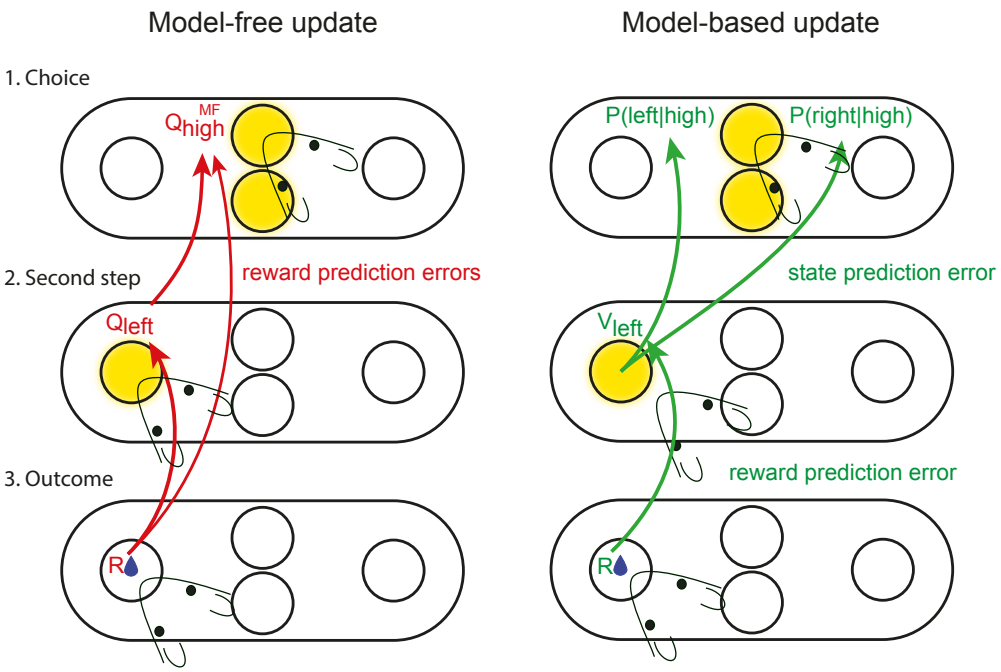
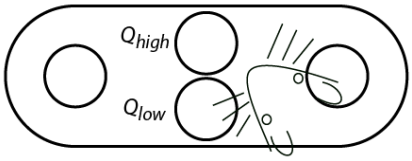


Figure 20: Model-free and model-based learning during a trial. In the above example, the agent chose a high poke, the left side lit up and the agent was rewarded. The model-free component (left) learns action-values, *Q*, directly using reward prediction errors (RPEs), that occur when the subject makes the 1st step choice and reaches one or other 2nd step state (each of which has an estimated value associated with it), and again following the 2nd step action which obtains or not a reward. In contrast, the model-based component (right) learns a model of which state follows which action, updating these estimates using state prediction errors, i.e. the difference between the state prediction and experienced next state. The state predictions are combined with estimates of the state values, *V*, (learned through RPEs) to give estimates of the action values at the 1st step. Image from (Akam T. 2016)

The agent employing both strategies learns separate model-free and model-based estimates of the action values and then combines these in a weighted sum to give net action values, which drive choice (Figure 21). Note that the weight for the model-free and model-based values are independent and not reciprocal, so that a subject that has a higher MB weight does not necessarily have to have an absolute lower MF weight.



$$Q_{high} = W^{MF} Q^{MF}_{high} + W^{MB} Q^{MB}_{high}$$

$$Q_{low} = W^{MF} Q^{MF}_{low} + W^{MB} Q^{MB}_{low}$$

Figure 21: The estimates of the action-values, *Q*, for choosing high or low are weighted sums of the action values based on model-free and model-based learning. Note, that the weights, *W*, of the model-free and model-based action-values are independent of one another (and not reciprocal).

Several different models developed by Akam (Akam T. 2016) from the original Daw two-step model (Daw et al. 2011) were considered. The model was selected by model comparison (see page 48). The 10 main parameters are briefly presented here:

- *alpQ*: Action-value learning rate for the action chosen
- *decQ*: Action-value forgetting rate (decay → zero) of the action not chosen
- *lbd:* Eligibility trace (i.e. temporary record) of the occurrence of the action
- *act:* Motor model-free weight
- *alpT:* Action-state learning rate for the action chosen
- *decT:* Action-state forgetting rate (decay →zero) of the action not chosen
- *alpCK:* Choice kernel learning rate
- *apSK:* Motor choice kernel learning rate
- *G_td:* Model-free weight
- *G_mb:* Model-based weight

The influence of the model-based and model-free components of the agent on behaviour are determined both by the weights *G_mb* and *G_td* given to their respective action value estimates, but also by the learning and forgetting rates in these systems which influence how much the action value estimates in each system fluctuate in response to experienced events. We, therefore, constructed a metric termed *influence* which combines these different factors to give an overall measure of the influence of each system on choice. This measure is defined for each system as the mean (over trials) of the absolute difference between the action value estimate for the low and high poke for that system. We calculate this metric for the model-based and model-free system, and large values indicate that the respective system has a strong influence on choices taken by the agent. Figure 22A shows the model fit and Figure 22B the corresponding influence of model-free

and model-based strategy on choice behaviour (Figure 22B). As we are primarily interested in how manipulations affect the influence of model-based and model-free control, only the mean absolute preference i.e. the influence of each RL system, will be presented in subsequent results. However, when comparing two populations or conditions, both plots and *P*-values from permutation tests can be found in the appendix.

The model also included four additional parameters:

- *bs*: Biased preference for high or low
- *sk*: Rotational bias (clockwise or counter clockwise)
- *ck*: Tendency of a previous choice to be repeated independently of outcome
- *ec*: *ck*, but with exponentially rather than linear decay
- *sc*: Motor learning choice kernel

The parameters *bs* and *sk* capture bias towards high/low and rational bias as in the logistic regression (page 41), while the parameters, *ec* and *sc*, capture choice perseverance (i.e. the tendency to repeat choices independent of outcome).

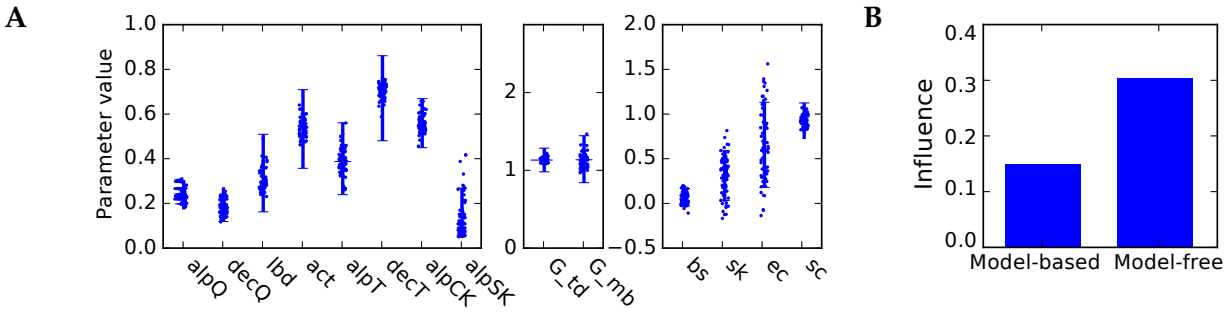


Figure 22: Example of the reinforcement learning model. A) The 14 parameters of the model fit from a portion of data from Water.1. Each dot is a session from a subject. B) Mean absolute preference for model-based and model-free strategy. Fitting of the subject data was done by using a hierarchical model, assuming random effects, such that each subject's parameters were drawn from a common population distribution.

2.3.6. Model comparison

To what extent does the RL model best describe the data and to what extent do the data support the model? Informed by previous model fitting on the water-deprived mouse data with numerous models with different parameters and degrees of complexity, we similarly compared a representative seven models on the food-deprived mouse data to identify the best fitting model. The desired model would also be used for the transgenic COMT mice, and both mouse types under the condition of acute stress, hence the model was studied across groups.

2.3.6.1. Method of model comparison

The Bayesian Information Criterion, BIC, was used to evaluate the quality of the fit of the models to the data, where the lowest BIC score is best (Schwarz 1978). A complex model will fit data better than a simple model by capturing more noise: a problem called overfitting, Figure 23 (Hawkins 2004). BIC penalises models based on the number of parameters to avoid overfitting.

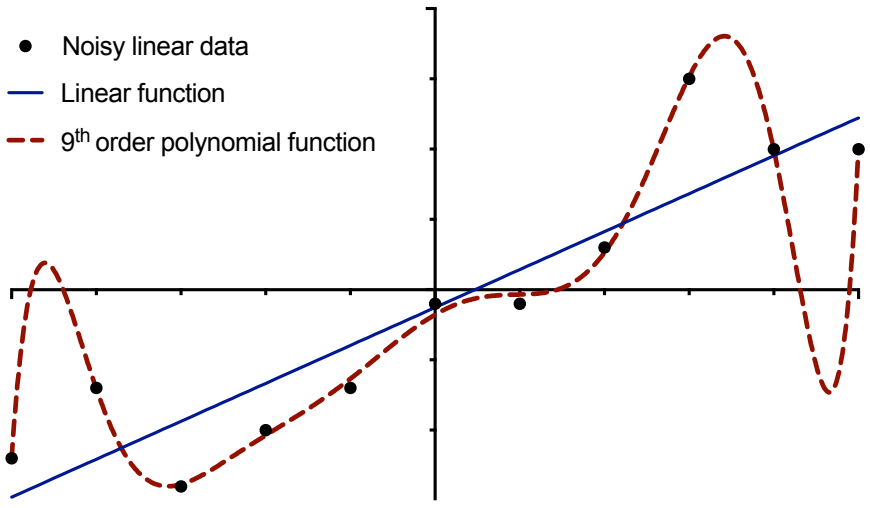
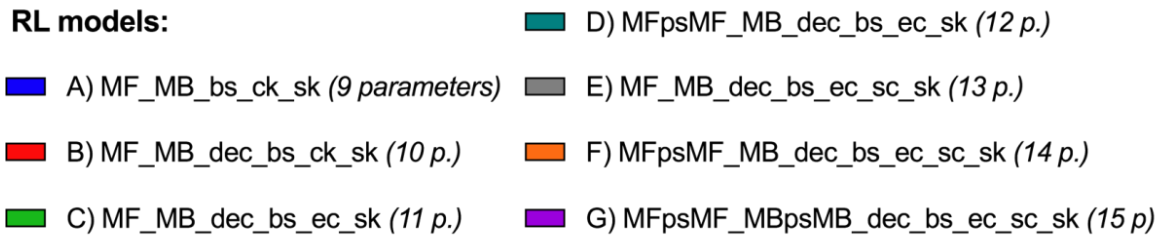


Figure 23: Overfitting is when a model describes random errors or noise instead of the underlying relationships. This can occur when a model is unreasonably complex e.g. having too many parameters compared to the number of observations, as seen above. Here 11 noisy observations of roughly linear data were fitted to a linear or a 9th order polynomial function. The polynomial function is a perfect (over)fit (R^2=0.99) compared to the good linear fit (R^2=0.84). However, the polynomial model is over-fitted and has poor predictive performance and overreacts to minor fluctuations or noise in the data. The linear model can be expected to generalise better and better representation the underlying relationship. Image adapted from Wikimedia Commons by Ghiles (CC BY-SA 4.0).

7 different model variants (A-G) of the RL model presented in section 2.3.5 (page 43) were compared. The models were a subset of the models developed by T. Akam for the water-reinforced mice (Akam T. 2016).



The RL models were fitted to the two cohorts, WT (also referred to as SUCROSE 2 in chapter 3 and 4) and the two COMT genotypes, separately, and both COMT genotypes during stress under the two conditions: WT None (from 214 sessions from WT not given injections before testing), WT Injection (80 sessions from WT given vehicle injections), COMT-Met (168 sessions from COMT-Met not given injection), COMT-WT (168 sessions from COMT-WT mice not given injection), and COMT Injection (96 sessions from COMT-Met and -WT mice given saline).

For simplicity only the results from the WT None data set is presented in the text below as the relative BIC score (ΔBIC) between each model. The combined results are display in Figure 24, and the individual BIC score are in the Appendix.

2.3.6.2. Results of model comparison

When learning the action-value (*Q*), of the chosen option, after each trial *Q* is updated and past experience is discounted, such that the learning rate, *alpQ*, is also the forgetting rate for the chosen option. RL models typically assume that action values of options that are not chosen remain unchanged, i.e. no forgetting occurs for these options.

However, data model comparison demonstrated that the addition of action value and transition probability forgetting *not chosen* options, parameterised by the decay rates, *decQ and decT*, dramatically improved the goodness of fit (Akam T. 2016). This was also observed in this data,

demonstrated by a big drop in BIC score (WT None: Model *A*-*B* ΔBIC 2607, Figure 24). Addition of an action value forgetting rate has previously been shown to improve model-fitting in another task: a simple bandit task in rats (Ito and Doya 2009).

Forgetting the value of not chosen options promotes choice perseveration. Therefore, if subjects were, in fact, strongly perseverative, this could mistakenly be captured by the forgetting parameters (*decQ and decT*). The Daw-model included a perseveration parameter, (*ck*: choice kernel) for repeating the previous choice independent of outcome (as *Choice* in the logistic regression page 41). However, this cannot capture perseveration effects which persist for multiple trials (Akam T. 2016). I, therefore, also tested a model with exponential decaying (*ec*) rather than the single trial (*ck*) choice kernel and found that this further improved goodness of fit (WT None: Model *B*-*C* ΔBIC 372, Figure 24).

While the choice kernels (*ec* and *sk*) assume that perseveration occurs at the level of the decision between the high/low poke, it is possible that the tendency to perseverate is at a lower motor level (Akam T. 2016). In the task, a given 1st step choice (high/low) involves a different motor action depending on which side (left/right) the previous trial ended on (i.e. left→high, left→low, right→high, and right→low). Incorporating an exponentially decaying motor level side choice kernel (*sc*), where action values of high and low are different depending on the side of the previous trial (i.e. whether the subject is returning from left or right) further improved model fit (WT None: Model *C*-*E* ΔBIC 256, Figure 24), indicating that perseveration effects at the level of motor actions are a feature of our data.

Since there was a motor level component to perseverance this raised the question of whether RL also occurred at both levels (i.e. choice and motor levels of representation). Models where model-free action values were learned in parallel for each action representation (*MFpsMF*) both in terms of choice (high/low) and motor action (e.g. left→high) generally improved this fit (Model *C*-*D*). However, also having parallel motor-level model-based action values (*MBpsMB*) did not improve

fitting (WT None: Model *F*-*G* ΔBIC -89, Figure 24). The other conditions (injection and genotype) showed a similar pattern (Figure 24).

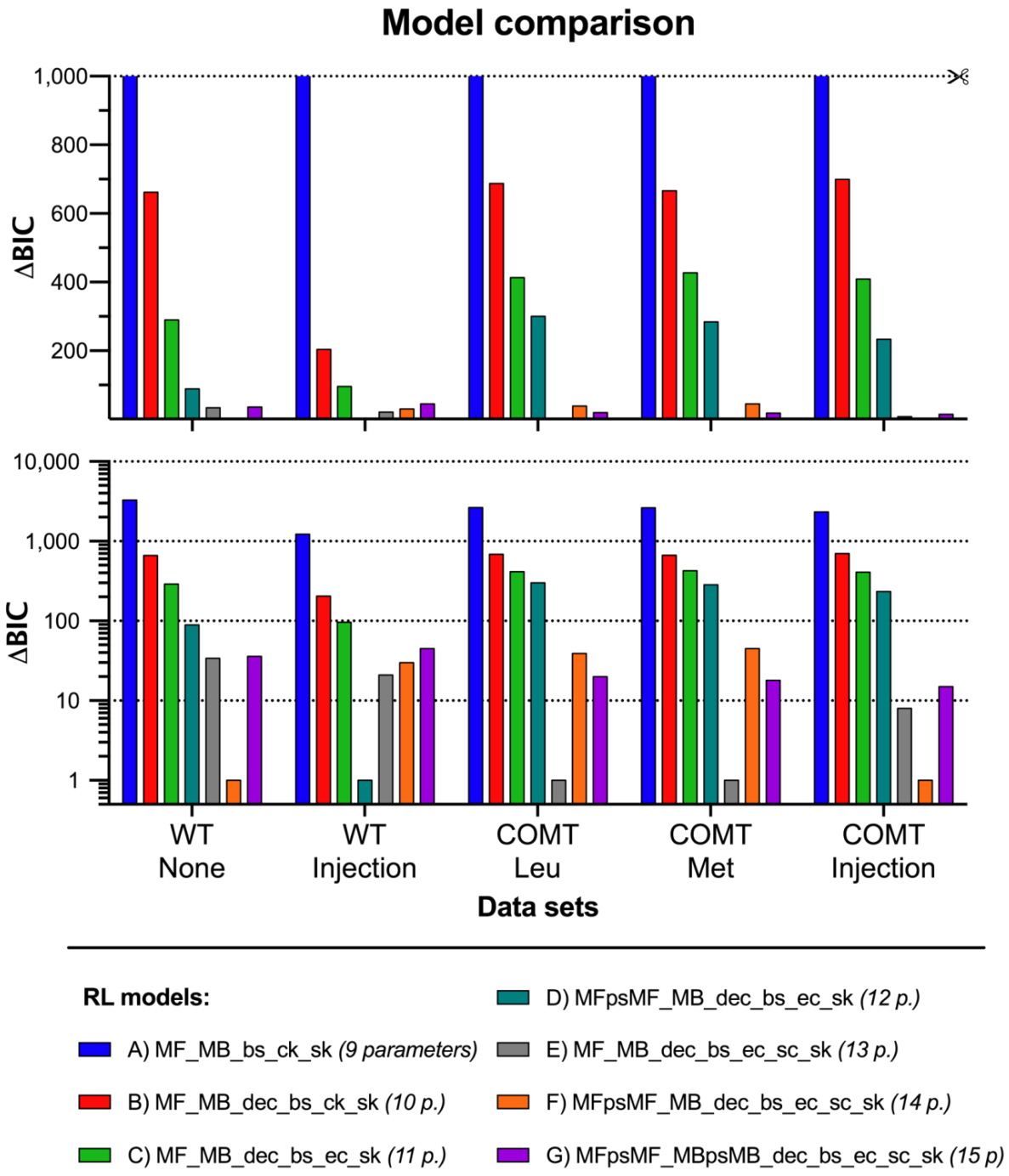


Figure 24: Model fit comparison of the data sets plotted as the differences in BIC score (lowest is best) compared to the best fitting model for the given data set. Displayed in A) on linear scale cut off at 1000 and B) on a logarithmic scale (the best model has been assigned ΔBIC=1 instead of zero for the purpose of visualization). The model comparison showed several well-fitting models. Due to relative low BIC scores (ΔBIC<100) among them, it was difficult to identify the best fitting model. WT None (from 214 sessions), WT Injection (80 sessions), COMT Met (168 sessions), COMT WT (168 sessions), and COMT Injection (96 sessions from COMT-Met and -WT mice).

This pointed towards three models with 13, 14 and 15 parameters with ΔBIC scores <100. This was too low-resolution to properly identify the best model. All models showed the overall same direction of loading on the main parameters (*alpQ, decQ… G_mb*), and the influence of model-free and model-based RL.

After much consideration, model *F* with 14 parameters was selected as it was overall the best model across genotype and conditions. Moreover, model *F* was also the best fitting model for the water-reinforced mice and the same conclusions were reached for the effect of the additional parameters and improvements (e.g. *dec* and *MFpsMF*) (Akam T. 2016).

Model *F* therefore allowed comparison across reinforcers, genotype, and conditions.

2.4. Permutation testing

Permutation testing is commonly used in large data sets such as genomic studies and fMRI studies (Phipson and Smyth 2010). The idea of permutation testing was first described by Fisher in 1935, with the assumption that *all* possible permutations would be computed to produce an *exact* P-value (Legendre and Legendre 1998). Apart from being easy to implement and apply, a great advantage is that permutation testing can be used regardless of sample distributions.

However, resampling and retesting is computationally demanding; even the small data set (n=20,20) given in Figure 25 yields 3×10^{29} possible permutations. We, therefore, limit the permutations to 1000 resamples giving an approximation to the true P-value. The permutations are randomly generated each time and the P-values, therefore, fluctuate somewhat when the test is repeated.

When testing for significant differences in a parameter between two conditions, e.g. Drug vs. Vehicle, the absolute size of the difference in that parameter is first evaluated for the dataset. The distribution for this difference under the null hypothesis (that the two conditions are the same) is evaluated by generating an ensemble of 1000 permuted datasets in which the assignment of data

to the two conditions is randomised (Figure 25). The *P*-value is given by the fraction of permuted datasets which have an absolute difference between conditions that is larger than the true dataset.

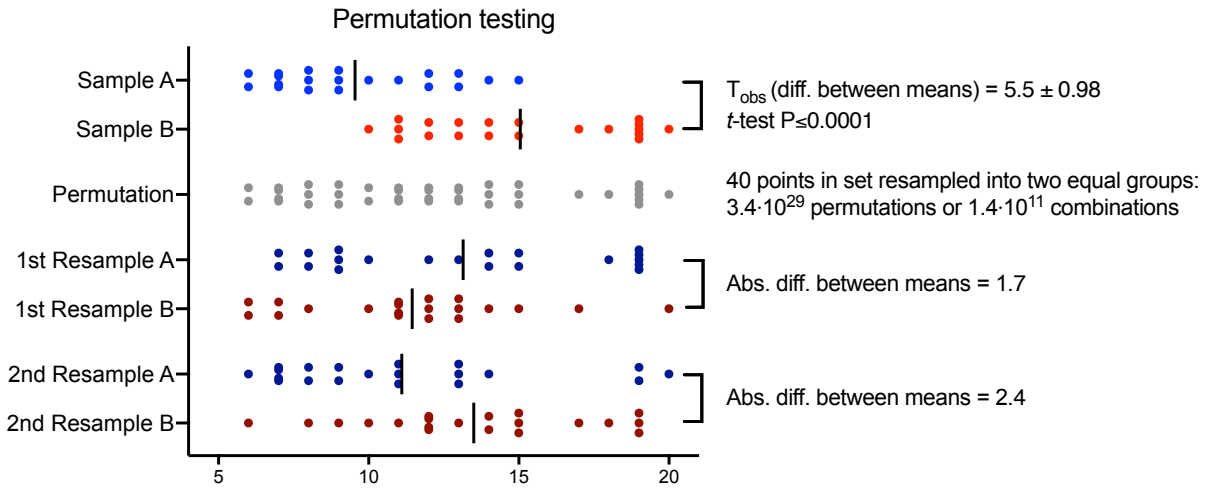


Figure 25: Permutation testing of two example groups labelled A and B. Initially, the observed difference in means between the two samples, $T_{obs.}$, is calculated. The two samples are then pooled and permuted (resampled) into new samples of corresponding sizes to the original groups. The difference in sample means is computed for every resampling. The set of these differences is the distribution of possible differences under the null hypothesis that group label does not matter. The two-sided P-value of the test is calculated as the proportion of sampled permutations where the absolute difference is $\geq T_{obs}$.

2.4.1. Permutation testing within subject and cycle

As the study was a within-subjects design, statistical comparisons and permutations were carried out within subjects. For permutation testing, this is done by permuting the condition labels on sessions within in each subject's data (i.e. a Vehicle session can be swapped with a Drug session from the same subject, but not any other subject). Thus, within-subject permutations, the assigned labels Drug and Vehicle are randomly rearranged within each subject.

Furthermore, the studies lasted a long time, e.g. the WT stress/drug study lasted 59 days, and each mouse experienced on average 30,000 trials and ~400 reversals during that period. As can be observed in Chapter 4, on page 84, a learning process occurs during this time. Therefore, a problem with permuting only within subjects would be swapping a session at the start of the study with a session and at the end of the study. The permutations were, therefore, further limited to occur within short time frames during which limited learning would occur, but including

enough sessions to allow for the conditions to be permuted within the subject. For the WT and COMT stress study this meant permuting within each injection cycle; for WT this was 11 days, and for COMT this was 6 days.

For the genotype comparisons, the relevant comparison is cross-subject, so permutations were done at the subject level by permuting the genotype labels on subjects, and thus swapping all sessions from a Met mouse with a WT mouse during permutation.

3. From water to sucrose reinforcement

3.1. Introduction

The two-step task allows for the dissociation of model-free and model-based reinforcement learning and can be used to study the effect of genotype (e.g. enzymes involved in neurotransmitter release or metabolism), environmental stimuli (e.g. stress) or pharmacological agents (e.g. neurotransmitter receptor agonists) on RL.

The aim of this thesis is to study the effect of functional genetic variation in the COMT gene in a mouse model, the effect of injection stress, and the interaction between these factors, on model-free and model-based reinforcement learning in a mouse model. This is in itself interesting, in light of the emerging literature on the interaction between stress and dopamine function on reinforcement learning (Abercrombie et al. 1989; A. F. Arnsten and Goldman-Rakic 1998; Dias-Ferreira et al. 2009; B. L. Murphy et al. 1996b; Olver et al. 2015). However, effects of injection stress are commonly overlooked in pharmacological studies of behaviour and can mask pharmacological effects, if the injection control (vehicle) is assumed to be the behavioural baseline (Drude et al. 2011). This important issue is the focus of chapter 4.

However, in preparation for the stress and genotype studies detailed in chapter 5, we needed to validate the task in our laboratory, particularly as we needed to implement a modification of the original protocol to use sucrose reward in food-restricted animals rather than water reward under water restriction (due to limitations of the current Home Office animal license).

3.1.1. Changing the protocol from water to sucrose reinforcement

The original rodent two-step task designed and validated by T. Akam and colleagues (Akam T. 2016) employed water as reinforcement. These mice only had access to water during the task and for an hour *ad libitum* during days without testing. Satiation during the 2-hour sessions was

reduced, by providing chow in the testing box, which maintained the animals' thirst during testing and also helped them maintain and increase their free-feeding weight.

Therefore, while we were still able to motivate restricted mice to work for reward, our first aim was to determine an optimal reward size of sucrose solution (i.e. concentration and volume) that would be big enough to lead to a high trial rate, but small enough to not lead to satiation during the session. Having adapted the protocol, the second aim (in a separate cohort) was to compare the differences in learning and task performance of water and sucrose-reinforcement in preparation for the study of injection stress in Chapter 4.

3.2. Methods

In our comparison, we include data from two C57BL/6 cohorts on water deprivation and water reward (WATER 1 and WATER 2) and two C57BL/6 cohorts on food deprivation with sucrose solution reward (SUCROSE 1 and SUCROSE 2). The data from the water-reinforced cohorts were provided by T. Akam and acquired at Fundação Champalimaud, Portugal, between 2014-2015. The sucrose-reinforcement studies were run in Oxford, as part of this project.

The weight (% of the free feeding weight) during the study was held at an average of 95% for WATER 1 and 2, and 85-90% for SUCROSE 1 and 2 (Appendix, page 141).

Cohort	**Size**	**Strain**	**Age (weeks)**		**Free Feeding Weight (g)**	
	n		*Mean*	*SD*	*mean*	*SD*
Water 1	6	C57BL/6	10.0	0.8	26.1	0.8
Water 2	6	C57BL/6	11.5	1.5	23.0	1.3
Sucrose 1	7	C57BL/6	10.9		25.2*	1.5
Sucrose 2	8	C57BL/6	12.3		21.1*	1.5

Table 4: Summary of baseline characteristics of water and sucrose-reinforced cohorts. *Due to the increasing age of the sucrose cohorts one gram was added to the free feeding weight after the initial week and another gram after the second week of testing to correct for growth. This was not done for the water cohorts.

3.2.1. Training

3.2.1.1. Water Cohorts

WATER 1 and 2 cohorts started with 10 µl water as the reward, which was stepwise lowered to 2 µl (6.5 µl day 5, 8 µl day 8, 2 µl day 21). Session length was always 2 hrs. The initial contrast of the probability of transition and reward contrasts (0.9/0.1) were lowered to (0.8/0.2) at day 12 (for transition) and 20 (for reward), respectively when they had reached an average of 500 trials and 5 reversals per session. A feature in the task software, *persistent variables,* enabled the transition and reward block information to be saved at the end of each session for each animal, ensuring they continued where they left off the next day.

In addition to the reward blocks (left / right side good), they also had a neutral reward block (reward probability for both options were equal: 0.4/0.4) and with a fixed length of 40 trials. The aim of the neutral blocks was to make it more likely the subjects track the reward probabilities as a continuous variable rather than inferring they are in one of two different states. Neutral blocks occurred pseudo-randomly on 20% of reversals.

The criterion for finishing training was a 3-day average of sufficient amount of trials: >450 trials/session, performance: >5 reversals/session, and reversal rate, τ: <25 trials. After training, they were tested until day 39.

3.2.1.2. Sucrose Cohorts

The sucrose-reinforced pilot cohort (SUCROSE 1) was trained according to the schedule used for water deprivation experiments, *including* neutral reward blocks. The cohort started with 10 µl 10% sucrose solution as the reward, which ended with 4 µl 20% sucrose solution. Session length started lasting 1.5 hrs and, by the end of testing, lasted 2.5 hrs. Transition contrast was lowered at day 21 after reaching an average of 477 trials and 6.8 reversals/session. Initially, the *persistent*

variables function was not enabled until day 15 and due to system failure was restarted again day 19 and 21. The task software feature, *persistent variables,* did not work correctly until day 21 of training.

The neutral blocks were removed from the training protocol at day 29. When they were removed from the pilot cohort there was no apparent effect on behaviour. Reward contrast was lowered on day 34 (average of 377 trials/session, 3.1 reversals/session). The reversal criterion was therefore lowered to >4 reversals/session. As τ is only calculated after non-neutral reversals, this criterion was unchanged (τ<25). Subsequent experiments did not include neutral blocks.

SUCROSE 1 was tested until day 36, followed by a drug study not presented here.

The second sucrose-reinforced cohort (SUCROSE 2) was initially made up of a total of 16 mice that were trained for 6 days. The most active 8 mice in the task at this point (highest cumulative activity) were selected for further training. The unselected mice remained housed in the cages so as not to disturb the established social structure and their data is not included.

They started with 10 µl 20% sucrose solution, which was lowered to 4 µl (Figure 26). Session length again started lasting 1.5 hrs and at the end of testing lasted 2.5 hrs. Transition and reward contrast were lowered at day 19 (average 220 trials and 2.2 reversals per session) and 24 (average 320 trials and 3 reversals per session), respectively. The *persistent variables* function functioned correctly. Mice in this cohort did not experience a neutral reward block.

In preparation for the study in Chapter 4 that started on day 39 without any injection or intervention, an intraperitoneal injection (i.p.) of 10 ml/kg saline was given on day 34.

3.2.2. Analysis

Task acquisition was measured by trials and the time in training days it took to reach the criterion for the full task, and sufficient trials and performance.

Task performance was compared unmatched for accumulated trials from day 30 when all cohorts were on the full task and had reached criterion and to day 35 before testing drugs (data not presented here), and also matched by average accumulated trials (WATER 1 days 28-30; WATER 2 days 27-29; SUCROSE 2 days 37-39)

The information on which block the subjects in SUCROSE 1 were in at the end of each session was not carried over to the next session due to the *persistent variables* function not working correctly. This meant an unintended block change happened at the start of each session, which was not the result of the subject properly reaching the performance threshold (moving average of 75% of correct 1st step choices); this also meant learning the current state (block) and the task became more difficult. Subsequently, SUCROSE 1 subjects were not properly trained on the task structure and were therefore excluded from performance analyses.

Sessions from SUCROSE 2 day 34 were excluded from analysis due to a mock injection in preparation for the subsequent study, and a session from three different subjects on day 30, 31, and 36 were removed due to a malfunction with the solenoids delivering rewards. A subject from WATER 2 was excluded from analysis due to a substantial rotational bias ('always' poking top after returning from the left nose poke, and bottom after returning from the right nose poke) defined as an outlying high loading on the rotation bias parameter of the logistic regression and similar parameter of RL model, together with an analysis nor described here (personal communication T. Akam).

A repeated measures analysis of variance (RM-ANOVA) was used to examine trial rate dependent on reinforcement. Unpaired *t*-tests were used to compare accumulated trials across cohorts. All other all statistical comparisons are by between-subjects permutation-testing (1000 permutations per test).

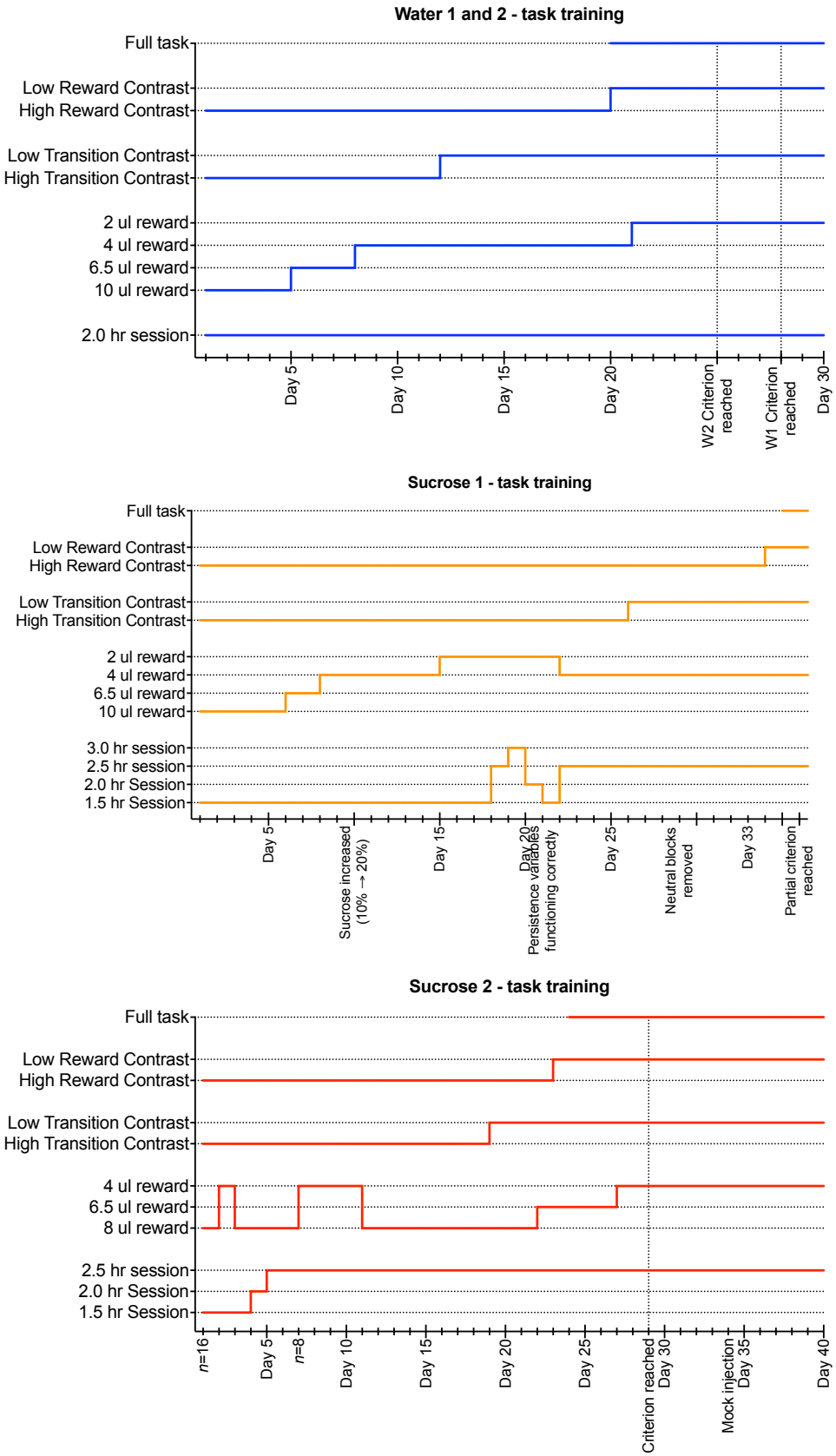


Figure 26: Summary of training of cohorts. Water.1 and Water.2 were trained on an identical schedule (blue). The criterion (dashed line) for having finished training was being on the full task (low transition and reward contrast), lowest reward size possible, and sufficient amount of trials and performance.

3.3. Results

3.3.1. Task training and trial rate

Sucrose mice compared to those with water as reinforcement showed a slower trial rate and thus slower trial accumulation during training (Figure 27). This was despite the fact that the lower trial rate had been attempted to be countered by extending the session length to 2.5 hours.

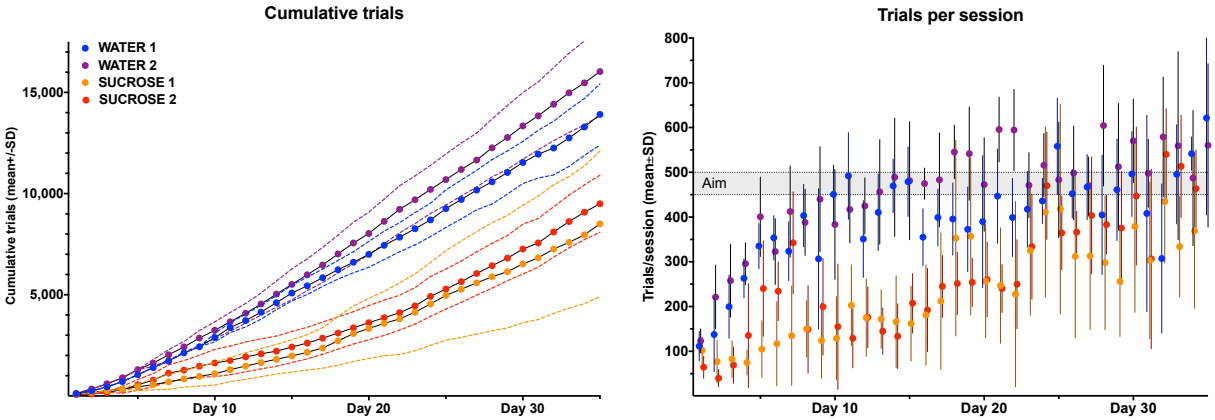


Figure 27: Average cumulative trials ±SD (left) and average trials per session±SD (right) during the first 35 days. The aim for the WATER 1+2 was 500 trials/session and 450 trials/session for SUCROSE 1+2.

During the first 35 days of testing (before any interventions) mice on water-reinforcement (1 and 2) experienced a mean total of 14,972 trials, while the sucrose-reinforced (1 and 2) experienced only 8,920 trials (Table 5). That is, there were 40% fewer cumulative trials using sucrose compared to water-reinforcement (unpaired *t*-test: t(25)=14, *P*<0.0001).

In an attempt to acquire a fully trained cohort more quickly for experimental testing, the SUCROSE 2 cohort was initially comprised of 16 subjects, where the most active 8 subjects were selected after day 6. Despite this selection, SUCROSE 2 did not complete more trials than SUCROSE 1, further demonstrating the effect of reinforcement type on trial activity.

Cumulative trials at end of day 35			
	mean	±	*SD*
WATER 1	13913	±	1511
WATER 2	16030	±	1289
Water average	14972	±	1497
SUCROSE 1	8383	±	3549
SUCROSE 2	9457	±	1481
Sucrose average	8920	±	759

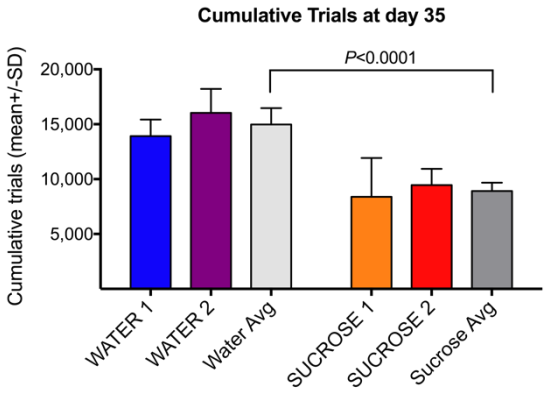


Table 5 and Figure 28: Comparison of cumulative trials between water and sucrose reinforcement at day 35. Water and Sucrose average by unpaired t-test: $t(25)=14$, $P<0.0001$.

The sucrose-reinforced mice accumulated trials at a reduced rate, due to the water-reinforced subjects having a higher trial rate and faster-accumulated experience with the task structure. During the start of the session, water deprived mice have a trial rate of 10 trials/min that gradually declines during the session and plateaus at around an hour to approximately 1.5-3.0 trials/min, ending at 2 trials/min. In comparison, the sucrose-reinforced mice start with a trial rate half that of the water-reinforced rate (5 trials/min) and plateau already after 20 min, to a stable rate of 2 trials/min (Figure 29). A RM-ANOVA of the trial rate during the initial 10 min and last 20 min yielded a significant interaction between time and reinforcement ($F(1, 23)=113$, $P<0.001$)

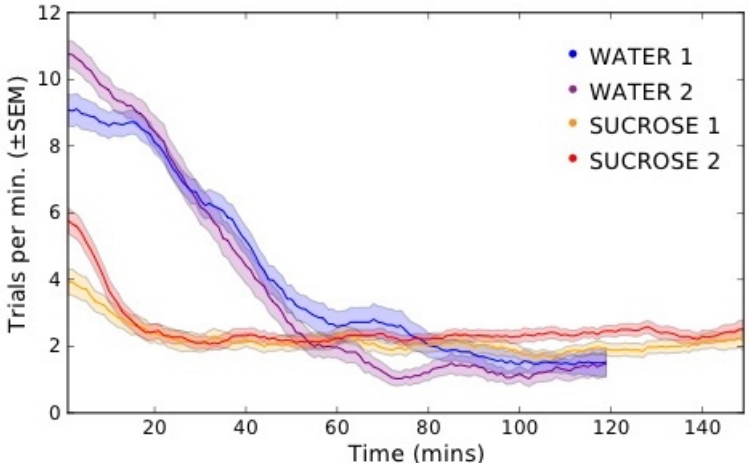


Figure 29: Plot of trial rate (trials per minute) during the time course of the session for water-reinforced during sessions day 30-35. The initial trial rate was 10.8, 9.1, 3.9, and 5.7 trials per min, and the number of trials per session was 450±22, 488±23, 328±29 and 374±17, for WATER 1, WATER 2, SUCROSE 1 and SUCROSE 2, respectively. Traces are trials per min ± SEM.

The reversal rate was compared between the water (WATER 1 and 2) and sucrose-reinforced (SUCROSE 2) subjects during training. Despite the differences in experienced trials (Figure 27),

both groups reached the combined reversal criteria of τ<25 after roughly the same number of days training: day 25 for WATER 1, day 28 for WATER 2, and day 29 for SUCROSE 2 (Figure 30).

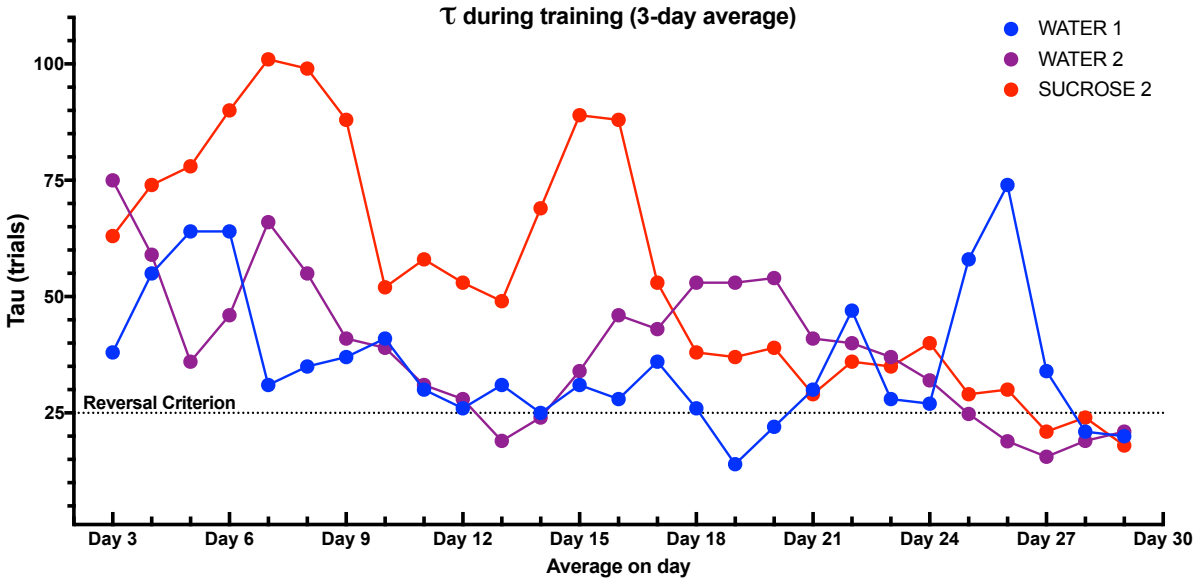


Figure 30: 3-day average of combined reversal τ during training. The combined reversal τ fluctuated during the early part of training and the difference in trajectory is likely due to lower the transition and reward contrasts on different days: day 12 and 20 for WATER 1 and 2 and day 19 and 24 for SUCROSE 2, respectively. The combined reversal τ criterion was reached on day 25 for WATER 1, day 28 for WATER 2 and day 29 for SUCROSE 2.

The WATER 1 and 2 reached training criterion (a 3-day average of sufficient amount of trials: >450 trials/session, performance: >5 reversals/session, and reversal rate, τ: <25 trials) at day 28 and 25, respectively. Due to a changes in task structure the training criterion was lowered to a 3-day average of >40 trials/session, >4 reversals/session, and τ: <25 trials). SUCROSE 1 only partially reached the new training criteria at the end of the study on day 36 with a 3-day average of 465 trials and 3.8 reversals per session, but τ = 35 trials. Failure to reach τ<25 trial within the study period was likely due to the *persistence variables* function not functioning correctly until day 21, which is why this cohort was not studied further. SUCROSE 2 reached training criteria on day 29 with a 3-day average of 398 trials/session, 4.2 reversals/session, and τ of 19 trials.

WATER 1, WATER 2 and SUCROSE 2, therefore, reached their training criteria at roughly the same time: days 28, 25 and 29, respectively.

In summary, to adapt the protocol from water to reinforcement by lowering food-deprivation from 95% to 85-90% of free-feeding weight, replacing a 2 µl water reward with 4 µl sucrose solution (20%) as the reward, and extending session length from 2 to 2.5 hours (Table 6), was required.

Protocol	**Food dep.**	**Cohorts**	**Reward volume**		**Sucrose concentration**		**Session length**		**Neutral blocks**
	FFW%	*Training:*	*start*	*end*	*start*	*end*	*start*	*end*	
Original	95%	WATER 1 and 2	10 µl → 2 µl		0%		2 hrs		Present
Pilot	85-90%	SUCROSE 1	10 µl → 4 µl		10% → 20%		1.5 → 2.5 hrs		Removed day 29
Final	85-90%	SUCROSE 2	10 µl → 4 µl		20%		1.5 → 2.5 hrs		Not present

Table 6: Summary of changes to the training protocol from the original water protocol to the pilot (SUCROSE 1) and final sucrose (SUCROSE 2) protocol. FFW%: % of free-feeding weight.

3.3.2. No performance differences between water and sucrose

After both groups had reached criteria for testing the water and sucrose-reinforcement subjects (WATER 1 and 2 pooled, n=13, and SUCROSE 2, n=8) were compared on reversal performance during day 30-35 calculated across the groups and statistical comparison was by cross subject permutation-testing (Figure 31, top).

The two groups were not significantly different in terms of the number of trials it took to reverse after block changes, which was independent of type (reward or transition block change): reversal τ after reward block changes was 36.6 (water) and 25.6 (sucrose) trials (calculated across the groups; cross subject permutation-testing: P=0.425), reversal τ after transition block changes was 22.1 (water) and 21.7 (sucrose) trials (P=0.937), combined reversal τ after all block changes was 29.7 (water) and 24.2 (sucrose) trials (P=0.554).

However, there was a trend-level significant difference in block end probability: 0.67 (water) and 0.63 (sucrose) (P=0.060), which indicated that WATER 1 and 2 were choosing the correct option more at the end of blocks and, hence, performing better than SUCROSE 2. However, this could be an effect of the water cohorts having accumulated more experience through having

completed ~50% more accumulated trials than SUCROSE 2.

The three cohorts were therefore *matched* and compared during days with the same average accumulated trials (Figure 31, bottom). When matched, there was again no difference regarding reversal performance: reversal τ after reward block changes was 23.6 (water) and 19.4 (sucrose) trials (P=0.663), reversal τ after transition block changes was 17.8 (water) and 16.7 (sucrose) trials (P=0.894), combined reversal τ after all block changes was 20.5 (water) and 18.7 (sucrose) trials (P=0.789). But now block end probability – 0.635 (water) and 0.634 (sucrose) – was also the same (P=0.980), suggesting that the previous difference was due to the difference in experienced trials (see Chapter 4, for detailed analysis of the effect).

In summary, while water restricted animals performed more trials and therefore advanced faster through training, there were no overall differences in performance with the two reinforcers.

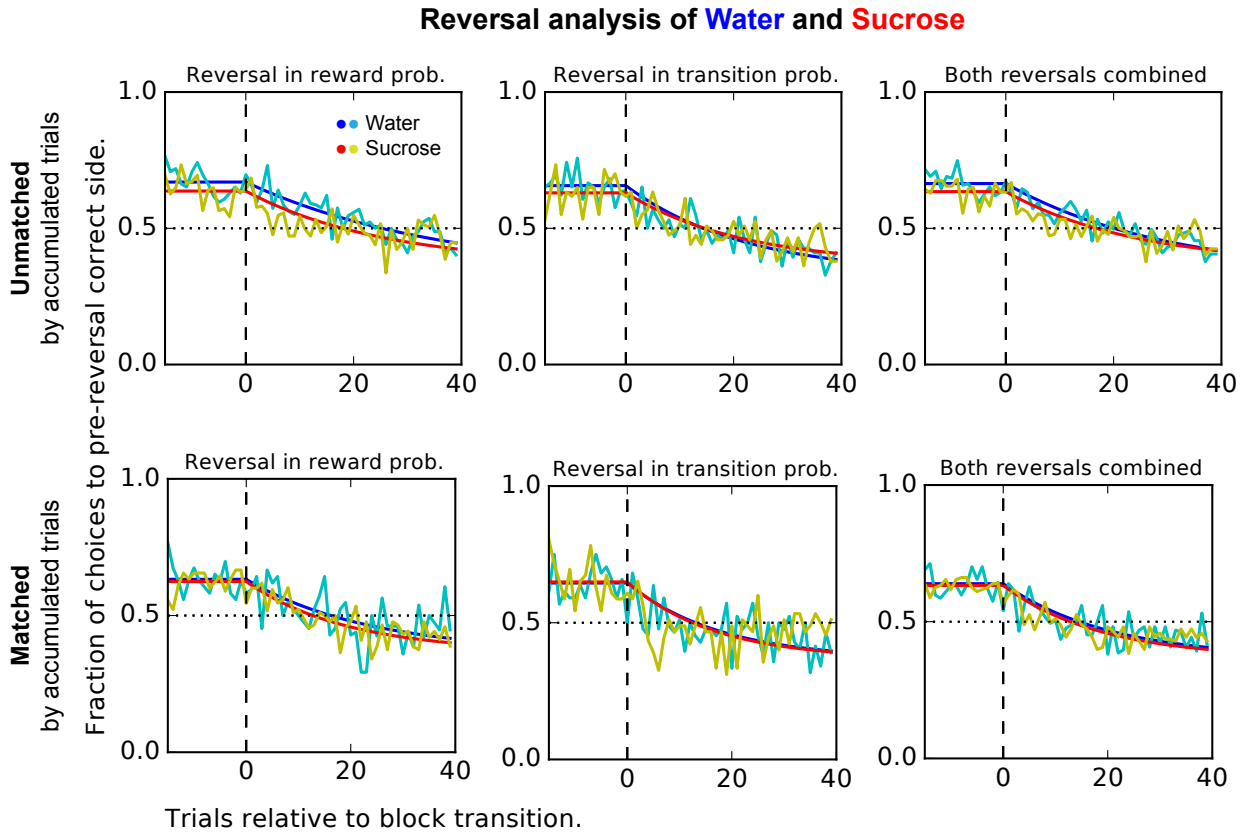


Figure 31: Reversal analysis for water-reinforcement with neutral blocks (WATER 1+2 pooled) and sucrose-reinforcement without neutral blocks (SUCROSE 2) unmatched and matched for experience (accumulated trials).

3.3.3. Difference in reinforcement learning

Model-free RL is a more efficient computation based solely on RPEs, while model-based reinforcement learning is a more demanding computation, based on predictions of values and states. Because sucrose-reinforcement led to a slower trial rate, this could lead to a lower influence of model-free learning and a greater model-based influence. The 2nd step reaction time during trials was therefore examined, followed by analysis of choice behaviour by logistic regression and the RL model.

A logistic regression model was used to determine differences in performance i.e. repeating *correct* choices, and whether one of the groups was more influenced by *outcome* – the tendency to repeat rewarded choices – or *transition* – choices after a common transition. The regression plots show loading on the different predictors for repeating a choice, and regressions were compared by cross-subject permutations. There was a significantly lower loading on *choice* for the sucrose-reinforced (P=0.020) and *transition-outcome* interaction (P=0.009, Figure 32 top), meaning that sucrose- compared to water-reinforced had a lower tendency to repeat choices regardless of the previous trial events (i.e. reward or transition). Yet, when matched for experience (accumulated trials) these differences were not present: choice (P=0.886) and transition-outcome interaction (P=0.300, Figure 32 bottom). There was no difference in the loading on *correct* choices unmatched or matched by experience. This demonstrated that there was not a difference between the water and sucrose on the influence by outcome or transition of the previous trial, nor did one group perform better than the other. For both groups there was little loading on outcome, transition and transition-outcome interaction, indicating that model-free and model-based RL is still low; however, this increases after further sessions (as seen in the Chapter 4.3.1).

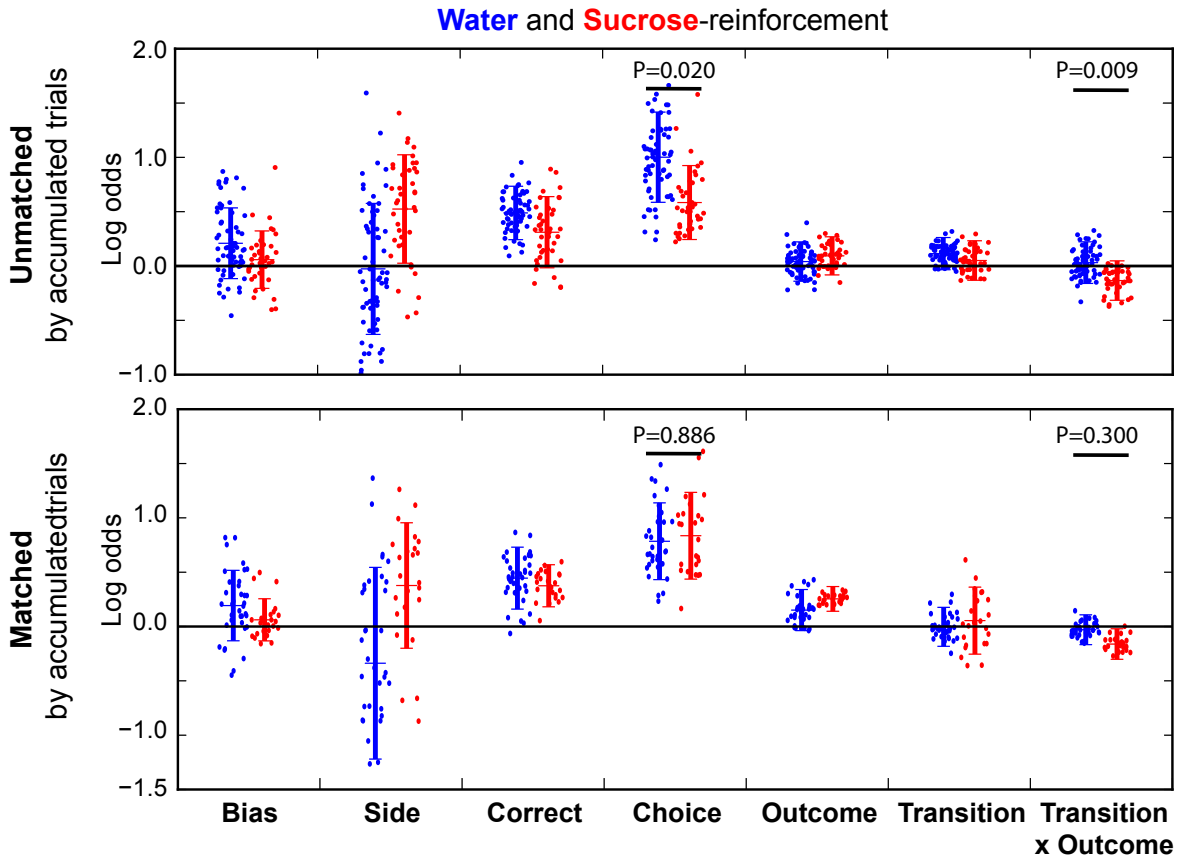


Figure 32: Logistic regression of choice behaviour using water (WATER 1+2 pooled, blue) or sucrose (SUCROSE 2, red) as reinforcement unmatched (top) and matched (bottom) for experience (accumulated trials). Unmatched, the sucrose-reinforced mice had a lower loading on Choice and Transition-Outcome interaction, which was not present after matching for accumulated trials. Cross-subject permutations.

To further study a possible difference between model-free and model-based learning, the water and sucrose groups were compared using the RL model and cross subject permutations (Figure 33) - unmatched and matched for experience (accumulated trials). The model showed that the water-reinforced subjects had up to a two-fold greater model-free influence on choice. However, in both comparisons, unmatched (P=0.094) and matched (P=0.067) for experience, this difference failed to reach significance (Figure 33). Similarly, there was no difference in model-based influence on RL either when unmatched (P=0.957) or matched (P=0.445) for experience.

This suggested that water-reinforcement, the presence of the neutral reward block, and the lower trial rate possibly led to a greater influence of model-free, but did not change model-based RL. Note that this comparison is necessarily limited in power by the small number of sessions

included because the subsequent study, since the SUCROSE 2 had to take precedence.

Nonetheless, it is possible to conclude, that the modified sucrose-protocol enabled the subjects to learn to perform the task at a similar level to water-reinforced subjects and sufficient for further experimentation.

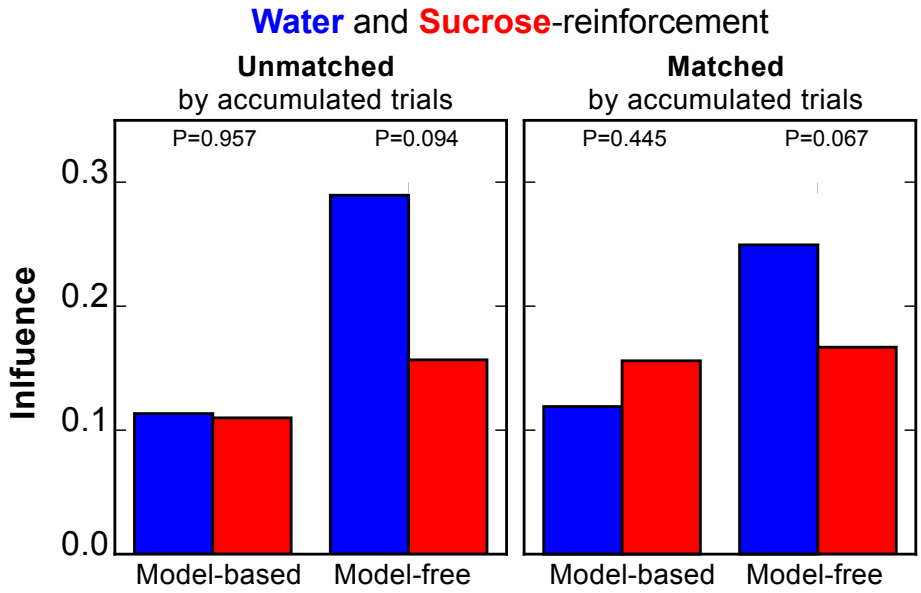


Figure 33: RL model of choice behaviour using water (WATER 1+2 pooled, blue) or sucrose (SUCROSE 2 , red) as reinforcement unmatched (left) and matched (right) for experience (accumulated trials). Unmatched, water and sucrose-reinforced subjects had equal model-based influence, but water had a two-fold greater, not significant, model-free influence on reinforcement learning. Matched, water and sucrose-reinforced subjects had similar model-based influence, but water had a greater, but not significant, model-free influence on reinforcement learning. Permutations across subjects.

3.4. Discussion

The original rodent two-step task was designed to dissociate model-free and model-based reinforcement learning employing water as reinforcement during water-deprivation. Here, the protocol was adapted to sucrose-reinforcement under food-deprivation.

The first aim was, therefore, to find a usable reward size – volume and concentration – and identify any additional task modifications needed for feasibility. In order to do this, the performance of wild-type mice was directly compared to those used in the original studies trained

in identical testing chambers but on water deprivation schedules.

Based on the pilot cohort (SUCROSE 1), we concluded that a 4 μl bolus of 20% sucrose solution was sufficient to be motivating, but not satiating during the session. Nonetheless, this still led to a lower trial rate compared to water reinforcement (Figure 29). This meant the mice completed fewer trials per session, thus slower training was seen, reducing the trials per session that could be used for analysis. To partially counteract this effect, the session length was extended from 2 to 2.5 hours.

A second modification was to remove the neutral reward blocks where left and right choices were equally valuable. This was originally included to make subjects more likely to track the reward probabilities as a continuous variable, rather than inferring them as in one of two different states. Since water-reinforcement also have neutral blocks, which are changes in reward probability, the water mice has extra reward prob. changes compared to transition prob. changes than the sucrose mice. This could contribute to the water cohorts having a higher model-free influence than SUCROSE 2. However, no difference was observed in performance (i.e. τ or block end prob.) after reward or transition probability reversals. The trials during the neutral block (40-120 trials per session or 10-20% of all trials per session) are not included in choice behaviour analysis, therefore, removing the neutral block also meant all of the trials in the session could be included in the analysis, which contributed to solving the problem of few trials per session. These modifications to the protocol and task are summarised above (Table 6).

Having the adapted the protocol, a second sucrose-reinforced cohort was used to compare the differences in training and task performance under water and sucrose-reinforcement. (Apart from the modifications to the protocol, it must be restated that the two groups were run under additionally different conditions (e.g. time, place, and handlers). Differences can therefore not solely be attributed to differences in reinforcement, but for simplicity are referred to as water and sucrose-reinforced below).

3.4.1. Trial rate and training time

For the sucrose-reinforced cohorts the trial rate was lower than the water-reinforced cohorts. For the pilot cohort this was likely aggravated by the non-functionality of the function, *persistent variables*, that saves the block information at the end of the session during the initial 21 days, which made the task more difficult (since the mice restarted each session in a new block).

We trained the second cohort (SUCROSE 2) using the optimised reward size and session length and selected the most active 8 of 16 subjects after six days of training. Nonetheless, this did not mitigate the lower trial rate compared to the water-reinforced cohorts, further demonstrating a likely effect of reinforcement protocol on trial activity. Future planned experiments using a comparable water deprivation protocol in Oxford to that employed in the original experiments will allow the contribution of other potential factors to be ruled out.

While the sucrose-reinforced cohorts did have an overall lower trial rate, these differences were manifested entirely in the first hour, where water-reinforced animals performed at very high rates. A high trial rate in a short period of time e.g. an hour as seen using the water-protocol may be useful for studying short-acting environmental, optogenetic or pharmacological effects. However, a lower, but stable trial rate that persists beyond 2 hrs as seen using the sucrose-protocol would be useful for studying longer lasting effects.

3.4.2. Performance differences

The performance of the second sucrose cohort was compared with that of the water-reinforced subjects. The slower accumulation of trials by the sucrose group could lead to a slower learning rate, as less experience was acquired. However, if a slower accumulation of trials leads to a greater informational weight/importance being put on or more experience extracted from each trial e.g. by prediction errors of greater magnitude, this could lead to a faster, if not equally fast, learning rate.

The performance differences were therefore analysed both as “unmatched”, i.e., performance on days 30-35 of testing in all animals, and “matched by previous experience”, i.e. the cohort average of accumulated trials, days between 28-39.

While there were some numerical differences between the water and sucrose-reinforced groups, there was no significant difference on reversals, and in logistic regression there was a lower tendency in the sucrose group to repeat choices, regardless of the previous trial events i.e. reward or transition (P=0.020), however, when matched for experience, even these subtle differences failed to reach significance. When the potential influence of model-based and model-free factors on behaviour was compared, the sucrose-reinforced had a lower, but non-significantly so, model-free influence. Both groups had similarly limited model-based influence (Figure 33), which was also observed in the low loading on transition and transition-outcome interaction on the logistic regression. However, both loading on the logistic regression and RL influence increases after further sessions as seen in the Chapter 4.3.1.

This suggests that differences in the water and sucrose protocols might have induced subtle differences in learning strategy. While gaining additional sessions could have improved this analysis, it was agreed that the priority was to move on to the pharmacological challenge experiment for SUCROSE 2, as no other comparisons were going to be made with the original water cohorts.

Nevertheless, it would be useful in future studies to resolve how different factors (e.g., reinforcement type, presence or absence of the neutral blocks, trial rates) influence learning. Several studies suggest that food or water deprivation enhances the incentive salience of cues paired with substances that satisfy a physiological need and thus make them effective reinforces (B. Balleine 1992). However, deprivation of any kind (i.e. water or food alone) may represent a high motivational state, as observed in a rodent study where water deprivation amplified Pavlovian conditioning and responding for sucrose reward, but sucrose concentration did not

impact behaviour (Tabbara et al. 2016). The two factors, reinforcer and neutral blocks, would have to be studied in two separate experiments or a 2x2 orthogonal design. To address the difference in trial rate and accumulated trials, which may be an effect of reinforcement type and could lead to different learning rates, subjects in the two groups could be *yoked* such that for a water subject the session ended, and no more trials could be initiated after having initiated as many as a specific sucrose subject. Alternatively, the intertrial interval could be set to dynamically be increased/decreased in a fashion leading to a similar trial rate.

In conclusion, in spite of some differences in performance when changing the protocol from water to sucrose reinforcement, the modified sucrose-protocol nonetheless enables the subjects to learn to perform the task at a similar level to water-reinforced subjects.

This in turn allows us to conduct further pharmacological and genetic experiments reported in the next chapters.

4. Acute stress on model-based/model-free RL

4.1. Introduction

The capacity to predict and control the consequences of actions is critical to being able to adapt successfully to changing environments; however, many decisions must be made under stressful circumstances. This chapter examines:

- How does acute stress influence decision-making strategies?

However, before I can study the effects of stress it is necessary to know how stable decision making strategies are over time. Therefore, this chapter first studies:

- To what extent do wild-type animals show changes in performance of the two-step task over time, both across and within sessions?

In order to investigate these questions, I therefore examined changes in task performance, both across and within sessions, and the effect of acute stress. These studies examined the performance of the previously-mentioned SUCROSE 2 cohort (chapter 4) on the two-step task during a 52-day long study.

4.1.1. Changes in behaviour over time

In model-based RL, the animal learns a model which predicts its future state given a chosen action, and evaluates possible choices by simulating their consequences several steps into the future. Such a system is slow and computationally demanding, but uses evidence efficiently, as the implications of changes in the environment can be evaluated in the model without having to experience them directly. In model-free RL, the animal learns only action values, which is done directly through reward prediction errors that reinforced rewarded actions. This allows for faster decision making, but excludes information about the task structure and experienced state

transitions.

Naturally, subjects improve their performance given more training on a task. For many learning processes the success criteria is the shift from explicit to implicit control on a behaviour, so-called automatisation, which frees cognitive capacity for other tasks. In RL, explicit or goal-directed control is the demanding model-based form of RL, while implicit or habit control is the faster, model-free form of RL. However, it is also possible that the subjects with more experience better understand the structure of the task and therefore become more model-based.

As observed in Chapter 4 subjects use a mixture of model-free and model-based decision making. When the study began, the subjects (SUCROSE 2) were well trained and had on average experienced ~11,000 trials so it might be expected that there would be few, if any, changes in learning within or across sessions subsequently. However, it is unclear how the learning process continues beyond the 39th session during repeated experience with the task during the following 52 days. This is potentially an important confounding factor for any study employing long sessions and a repeated-measures design as in Chapter 5.

I therefore investigated if task performance, and model-based and model-free, influence changed both *across* and *within* sessions.

Changes across sessions

During training in instrumental tasks behaviour may become more guided by the triggering stimulus and less independent of the outcome (habitual learning of stimulus–response) (Adams and Dickinson 1981; Wood and Rünger 2015). However, habit formation can be limited by complexity of the instrumental task, e.g. where choices of different actions lead to different rewards (Colwill and Rescorla 1985), meaning habit formation in the two-step task may be limited

In human studies using the two-step task, the participants only complete one or two sessions of 200-300 trials, and are thus not informative regarding changes across multiple sessions (Thomas

Akam et al. 2015; Daw et al. 2011; Doll et al. 2016; R. A. Otto et al. 2013d; Radenbach et al. 2015; Sharp et al. 2015; Voon et al. 2015b; Wunderlich et al. 2012). Moreover, the human task, with its two decision steps and moving reward probabilities of four options, is considered too complex for implicit learning, thus limiting across session changes (T. Akam, personal communication, June 2016). To date, however, session changes have not been studied in the rodent two-step, though the number of sessions and trials makes the data amenable to this approach.

As described in the General Methods, the rodent two-step task is a reduced version of the human two-step. It is a complex task, given that it has sequential actions and probabilistic reward delivery. Therefore, increased experience of the task and its structure could lead to a greater model-based influence as mice get increasing evidence of the task structure. Alternatively, as long sequences of choices require a fixed set of actions (e.g. nose poke up then left / down then right on common transitions) for the same type of reward (sucrose), repeated testing could lead to overtraining, and a shift from goal-directed to habitual responding (as seen by a greater influence of model-free learning).

I hypothesised that the mice would become more familiar and sensitive to the task structure with increased training, leading to a greater overall influence of model-based learning across sessions

Changes within sessions

The two-step is an appetitive task and, as seen in chapter 4 with both water and sucrose-reinforcement, there is an initial period of high trial activity, followed by a period of lower responding. In the water-reinforcement protocol, satiation during the 2-hour sessions was limited by providing chow in the testing box, which maintained the animals' thirst during testing. Despite a decline in trial rate during the session, they displayed stable behavioural strategies in terms of reinforcement learning (Akam T, personal communication, Dec., 2015 and Appendix 5.2.1). With

sucrose-reinforcement there was an initial period of high responding, followed by a lower, but more stable pattern of responding. This could be due to an initial acute hunger (which could be stressful). This is presumably followed by satiation; as a result, the value of the sucrose reward would be anticipated to decrease during the session. I therefore hypothesised that task performance would shift from model-free to more model-based *within* sessions.

4.1.2. Acute stress

Stressors are external stimuli or events that cause stress to an organism (i.e. heat, pain, restraint, noise, etc.) that result in a behavioural and cognitive response. This response-based stress approach holds that stress is defined by a pattern response that results from exposure to a given stressor. This definition is considered inadequate in human cognitive psychology, since it does not incorporate concepts of the individual's appraisal of stimuli or events in shaping their responses (Staal 2004; Stokes and Kite 2001; Tepas and Price 2001). However, in this study of acute stress in animals, I rely on the definition of an acute stressor as an uncontrollable adverse event and acute stress as the behavioural response elicited.

Both acute and chronic stress can increase reliance on habits (L. Schwabe and Wolf 2013). Model-based RL is computationally demanding: it relies on the PFC and taxes the central executive resources (B. W. Balleine and O'Doherty 2010).

In human two-step studies, depletion of working memory (WM) capacity, a core executive function, decreases model-based contribution to decision making (Gillan et al. 2015). Acute psychological stress (e.g. cold stress) reduces model-based contribution to decision making. However, this reduction is limited in individuals with large WM capacity (R. A. Otto et al. 2013d). A study using psychosocial stress (Trier Social Stress test) found that the reduction in model-based learning by acute stress depended on individual cortisol reactivity and chronic stress (as measured

by past stressful life experiences) (Radenbach et al. 2015).

However, tasks that rely on habitual functions of the basal ganglia are spared or even enhanced by stress (A. F. Arnsten and Goldman-Rakic 1998; B. L. Murphy et al. 1996b; B. L. Murphy et al. 1996a). This may explain the stress-related shift towards model-free learning. In summary, human studies of acute and chronic stress demonstrate a bias toward habit and model-free learning.

Animal studies of the effect of stress on goal-directed action have focused on the effects of repeated stressors. Two studies in rats showed that chronic administration of corticosterone, or exposure to multiple stressors (social defeat, forced swimming and restraint), led to shifting from goal-directed to habitual responding in instrumental tasks during outcome devaluation (Dias-Ferreira et al. 2009; Gourley et al. 2013). Similarly, a study exposing rats to multiple, intense, acute stressors (i.e. 20 min restraint, 20 min on an elevated platform, and 20 min restraint with five tail shocks, all in a brightly-lit room with loud music) rendered previously-acquired instrumental actions habitual in a two-lever choice task in extinction (Braun and Hauber 2013). However, in the same study, rats acutely exposed to a single stressor (60 min restraint), or a pharmacological increase in glucocorticoid and noradrenaline activity (by corticosterone and yohimbine injection) continued to display goal-directed responding that was sensitive to outcome devaluation.

Studies in rats support the notion that stress promotes habitual responding; however, there are currently no animal data on the effects of stress on performance of a multi-step decision-making task that can disassociate model-free and model-based learning. In this chapter, I therefore examine the effect of acute stress on the rodent two-step task. We hypothesised that acute stress in mice would, as in humans, lead to a decrease in model-based and an increase in model-free reinforcement learning.

4.1.2.1. Choice of stress model

There are several commonly-used rodent models of acute and chronic stress that can be divided into physical and psychosocial, or naturalistic, stressors (Figure 34). However, a commonly overlooked source of stress in pharmacological and behavioural studies is stress from an injection alone.

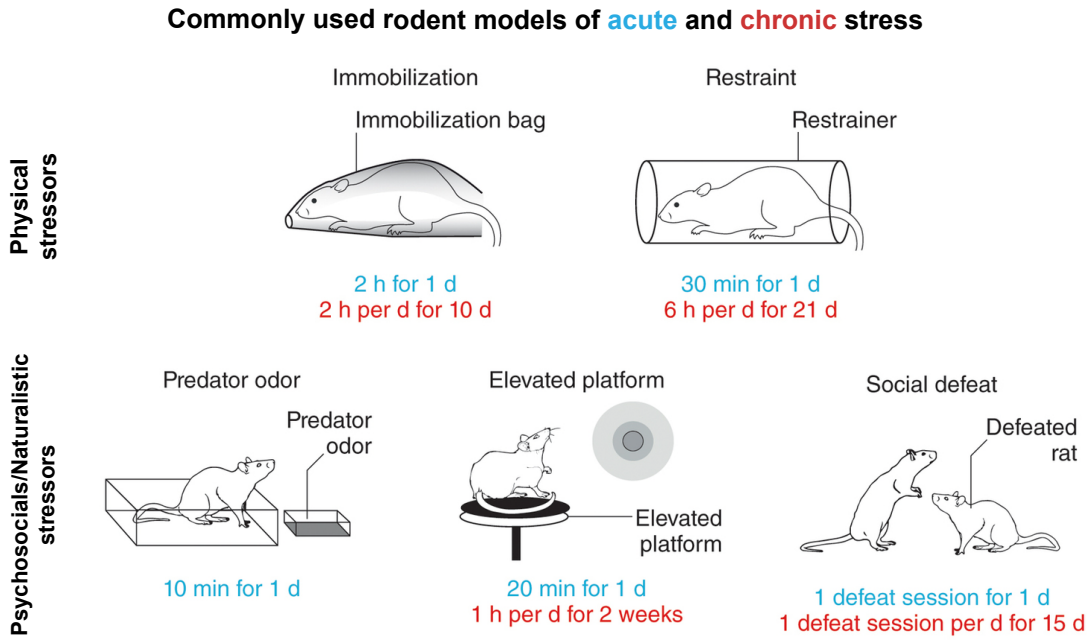


Figure 34: Common procedures used to study the effects of stress on behaviour and neural plasticity in rodents (acute stressors, cyan; chronic stressors, red). Immobilisation and restraint stress are widely used physical stressors. In contrast, naturalistic stressors are also used to trigger innate fear. Commonly used models of psychosocial stress include exposure to predator odour, a bright elevated platform, social defeat, and maternal separation (not shown). Adapted from (Chattarji et al. 2015)

In pharmacological studies where substances are administered by injection, researchers must give a non-toxic vehicle, such as saline or cyclodextrin, as the *control*. However, single or repeated injections of vehicle can alter baseline levels of glucocorticoids, heart rate, blood pressure and body temperature in mice (Drude et al. 2011; Meijer et al. 2006; Ryabinin et al. 1999). Therefore, such injections and the associated stress, may affect physiological parameters such as stress hormone concentrations and behaviour in control animals, potentially complicating study results (Balcombe et al. 2004; Drude et al. 2011). Indeed, we have seen this effect, i.e. a change from pre-injection

performance even with saline injections, and in a COMT -dependent manner (Barkus et al. 2016; Linda M. Laatikainen et al. 2013). However, we do not know what effect inject stress has on reinforcement learning e.g. does it change motivation or change performance?

We, therefore, asked if injection stress affected model-free and model-based learning by comparing sessions following an injection of vehicle, compared to sessions not preceded by an injection, from a pharmacological study conducted in the SUCROSE 2 cohort (which forms the focus of another student's thesis and so is not included here). Studies of stress employing the two-step task in humans have found a shift towards model-free learning (A. R. Otto et al. 2013b; Radenbach et al. 2015). We, therefore, hypothesised that injection stress in mice would similarly lead to an increased influence of model-free learning.

Rationale for study design

The physiological response to an acute stressor in mice is a quick rise in heart rate and plasma levels of glucocorticoids with a peak within approximately 15 min and a decline over 1-2 hours (Figure 35). In a male BALB/c mice plasma CORT returned to baseline within 2 hrs after a vehicle injection (Drude et al. 2011). We therefore analysed the initial 1.5 hrs of each session, in order to capture the peak of the CORT response to injection stress (Figure 35A), while also including a sufficient number of trials and reversals to permit analysis.

The pharmacological study had two drugs/vehicle injection time points: either 15 or 90 min before testing. These two time points allowed us to study whether the time since injection modulated the behavioural effects of the injection. According to the Inverted-U model of stress a low-intensity stressor can be beneficial to performance by increased arousal (Sapolsky 2015; Yerkes and Dodson 1908).

Thus, I hypothesised that the most recent injection (15 min before) would show the greatest changes in behaviour, compared to no injection. Furthermore, since the acute effects of stress

decrease with time, I predicted that the injection at 90 min would represent a lower level of stress, and thus show an intermediate change in behaviour, compared to the 15 min vehicle injection (Figure 35B).

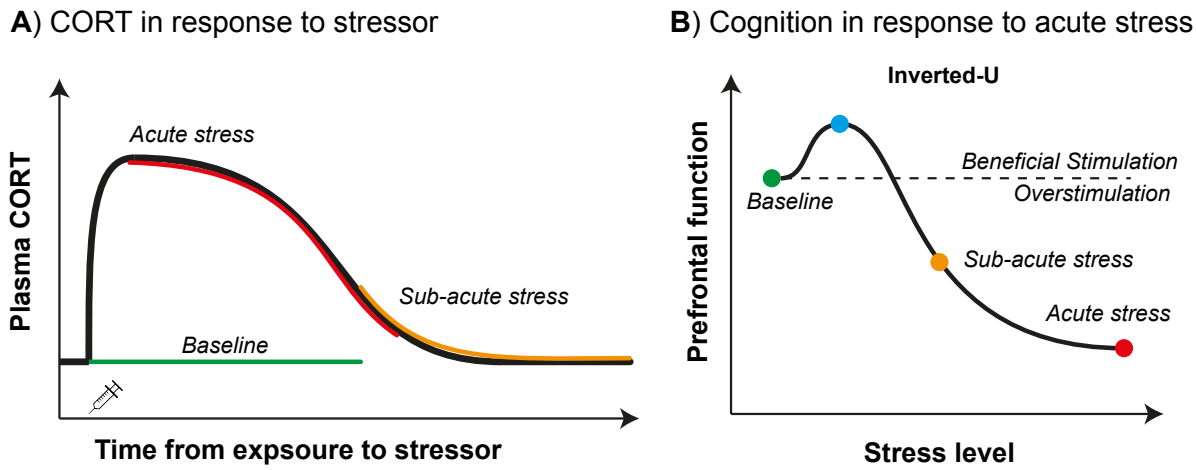


Figure 35: Glucocorticoid release and cognitive function in response to acute stress. A) Plasma CORT levels rise quickly due to stimulations of the HPA-axis and returns to baseline over a longer time course (curve extrapolated from (Drude et al. 2011)). Testing 15 min after injection would capture the acute stress response and testing 90 min after injection would capture sub-acute stress. B) The inverted U-model states that a low-intensity stressor can have a beneficial effect on cognitive function (Sapolsky 2015). However, I assume, that sessions 15 and 90 min after an injection will be above this threshold and be detrimental to task performance.

4.2. Introduction

4.2.1. Subjects and training

The cohort (SUCROSE 2) and its training is described in the previous chapter 4. In brief, 8 mice were trained on a 2-step task using 20% sucrose solution reinforcement (see Figure 36 for details). They finished training on day 29 when the group on average reached criteria with a 3-day average of 398 trials/session, 4.2 reversals/session, and τ of 19 trials.

A mock intraperitoneal injection (i.p.) of 10 ml/kg saline was given on day 34 (in preparation for the pharmacological study).

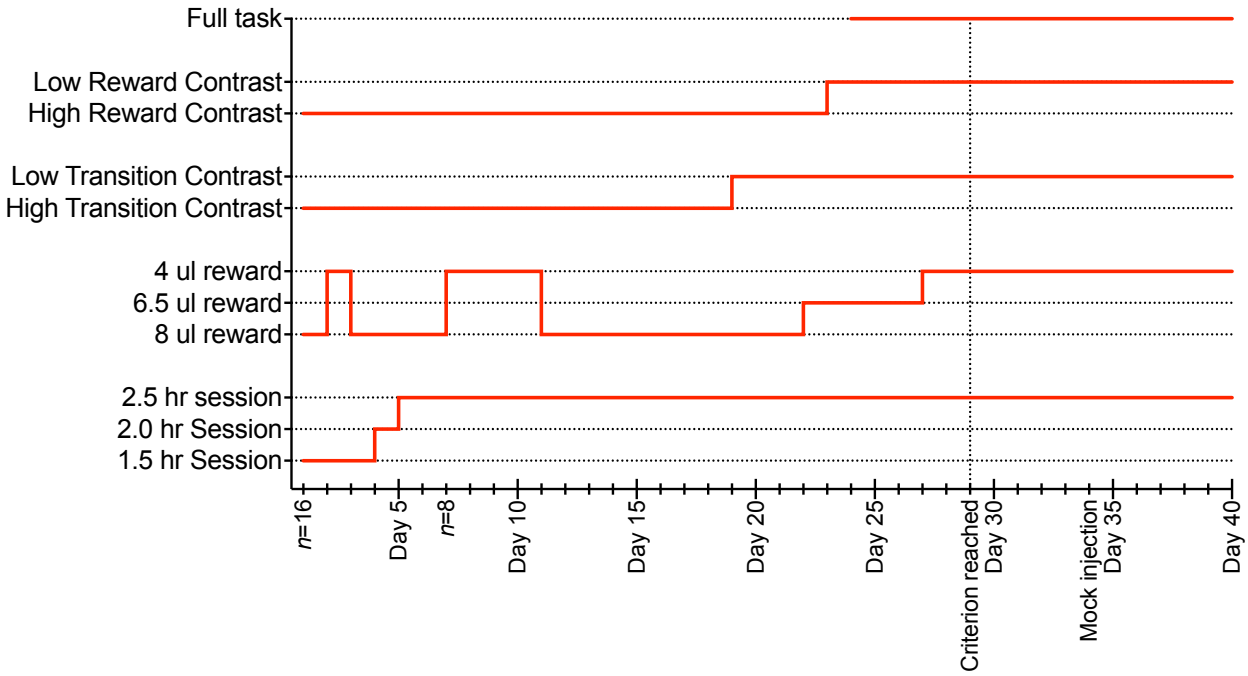


Figure 36: Summary of training of SUCROSE 2.

4.2.2. Behavioural study

The pharmacological study was initiated on day 39 as described in the General Methods. Briefly, the mice were given an i.p. injection with a drug or vehicle, 15 or 90 min before testing (the drug sessions are not reported in this thesis). Each mouse cycled through this regime 5 times (order counterbalanced across animals). The vehicle was 20% cyclodextrin in 0.9% (w/v) sterile saline. Injection days were followed by testing days on which no injection was given (Figure 37).

The testing sessions were (as the training sessions) 2.5 hrs in duration.

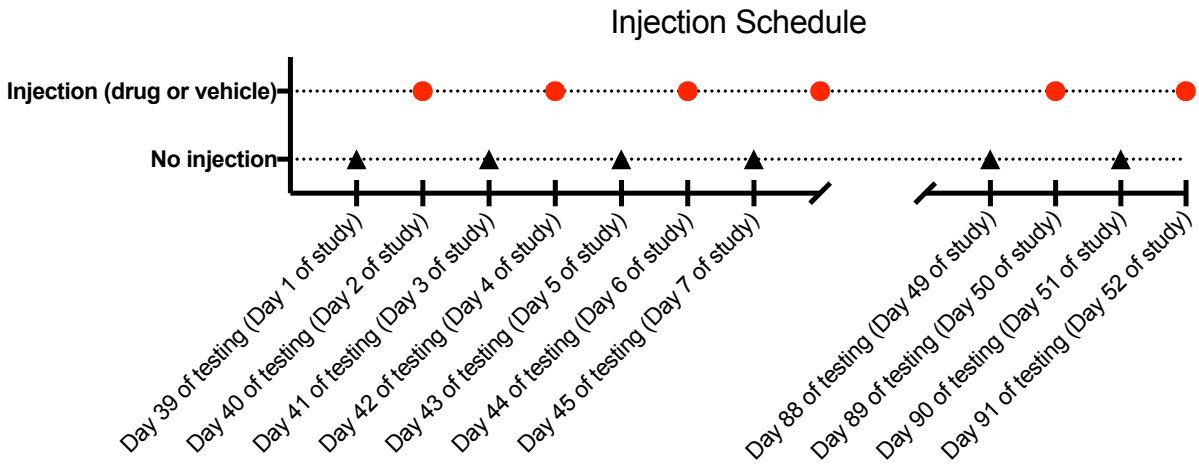


Figure 37: Injection schedule. The shift from odd to even 'no injection' days was due to a work limitation that led to a 2-day pause after day 71. An additional 'no injection' day (day 72) was added to resume the otherwise uninterrupted study.

The study lasted 52 days and a total of 424 sessions were collected, but five sessions were removed due to malfunction of the box or solenoids (Table 7).

Condition	Vehicle injection 15 min prior	Vehicle injection 90 min prior	No injection	Total
Sessions	38	40	214	419
Sessions/subject	4.75	5	26.8	52.3

Table 7: Distribution of sessions studied. The total count includes the two drug conditions, which are not studied here.

4.2.3. Analysis

Prior to analysing the injection stress data, we wanted to establish how stable performance was both across the experimental period and within each session. The first and last eight days without injections of the study were compared to study *across* session changes. As the days with and without injections alternated, the first and last eight days with and without injection actually spanned the first and last 16 days of the study. Statistical comparisons were made using within-subject permutations. To examine *within*-session changes, all the days of the study without injections were divided into three time bins: the first 30 min to capture the initial trial rate decline,

and the stable trial rate period was equally divided into the middle and last 60 mins of the 150 min sessions, which were compared to each other using within-subject and *cycle* permutations. Cycles are timeframes (11 days) in which sessions after an injection 15 and 90 min before occur. Limiting permutation within these time frames prevent an injection session at the end of the study from being swapped with a session with without an injection at the start, thus limiting across session changes (as explained on page 53 and observed on page 84). A*cross* and *within*-session changes were analysed with regards to trial rate, 2nd step reaction times, reversal analysis, and logistic regression and RL modelling of choice behaviour.

To study the effect of injection-stress, the sessions from three conditions were compared: following a vehicle injection 15 min before, a vehicle injection 90 min before, and no injection before. The effects were again studied by comparing trial rate, 2nd step reaction times, reversal analysis, and logistic regression and RL modelling of choice behaviour. In order to focus on acute stress effects from the injection, only the initial 90 min of the sessions were analysed. To control for the fact that there are changes across sessions statistical comparisons were made using within-subject and cycle (11 days) permutations, which essentially made each injection session paired with the time-wise adjunct five no injection sessions.

4.3. Results

4.3.1. Behavioural changes *across* sessions

The first and last 8 days of the 52-day long study were therefore compared to study *across* session changes, such as learning and habitual responding, to see if there was evidence for these changes in our study. Mean (±SEM) trials per session increased significantly from 403±13 to 531±25 (~33%) from the start to the end of the study (within-subject permutation test: $P<0.001$, Figure 38a). However, the overall pattern of responses remained the same: an initial high activity during the first 20 min followed by a lower plateau.

The 2nd step reaction time after common transitions also became shorter with increased training (within-subject permutations, $P<0.001$), but remained unchanged after a rare transition ($P=0.726$); consequently, the time difference between common and rare increased ($P=0.016$, Figure 38b).

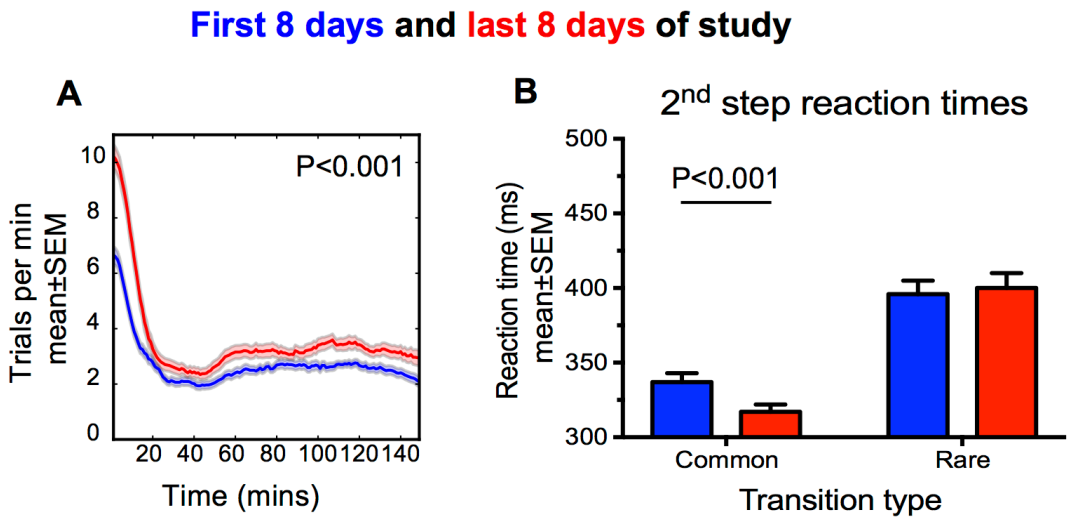


Figure 38: Trial rate and reaction times during the first and last eight days of the 52-day study. A) Mean (±SEM) trial rate increased from 403±13 to 531±25 trials (within-subject permutation-testing $P<0.001$). B) Mean (±SEM) 2nd step reaction times after common transition was ~20 ms shorter at the end than the beginning of the study (within-subject permutation-testing $P<0.001$).

We performed an analysis of reversals to identify performance changes across sessions: τ after reward reversals showed a numerical decrease from an average of 20.0 to 15.0 trials, although this change failed to reach significance (within-subject permutations, $P=0.114$), τ after transition

reversals decreased from 15.1 to 13.5 trials (P=0.690), and τ after both reversal types decreased from 18.4 to 14.6 (P=0.099) across sessions (Figure 39). The average block end probability of choosing the high reward option at the first step increased across sessions from 0.63 to 0.67 (within-subject permutations, P=0.018), which demonstrates that the subjects perform better at the end of the study compared to the beginning.

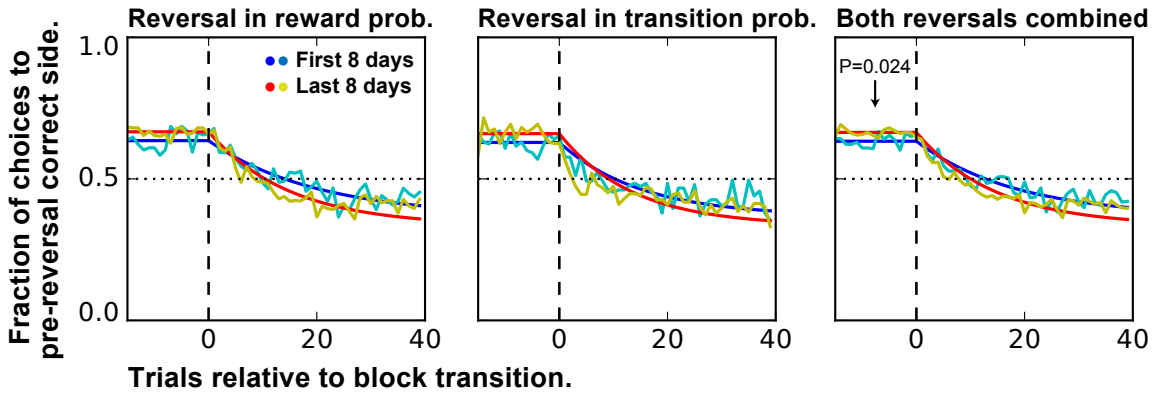


Figure 39: Reversal analysis of the first and last eight days of the 52-day study. τ after a reward reversal decreased from 20.0 to 15.0 trials, τ after transition reversal decreased from 15.1 to 13.5, τ after both reversal types decreased from 18.4 to 14.6, and average block end probability increased from 0.63 to 0.67 (within-subject permutation-testing, P=0.024).

The numerical, though not statistically significant, improvement on reversal τ combined with the significant increase in block end probability suggested a general improvement on task performance with increased training. We used logistic regression and RL model to further disassociate if the improvement was on model-based, model-free, or both learning strategies. The logistic regression showed an increased loading over the course of the study on the predictors: correct (within-subject permutation-testing, P=0.043), choice (P<0.001), outcome (P=0.002) and transition (P=0.022), but not transition-outcome interaction (P=0.918; Figure 40 left). The increase in both outcome *and* transition suggests an increase in model-based learning, despite no increase in interaction.

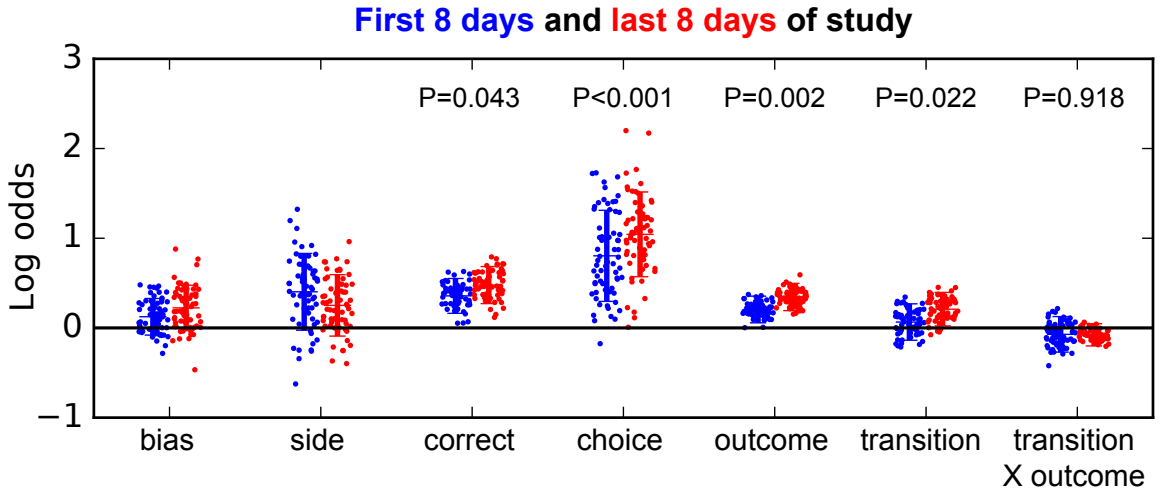


Figure 40: Logistic regression of the first and last days of the study. Left) The logistic regression showed an increased loading on correct, choice, outcome and transition, but not transition-outcome interaction, i.e. the subjects had an increase tendency to choose the correct 1st step, repeat the same choice independent of reward or transition, repeat the choice following a reward or a common transition, but not following a rewarded common transition.

This finding is supported by the RL model, which also showed that the only reliable increase was in the model-based influence across sessions (with-subject permutation-testing, P<0.001). While there was a minor increase in model-free influence, this was not significant (P=0.2, Figure 41).

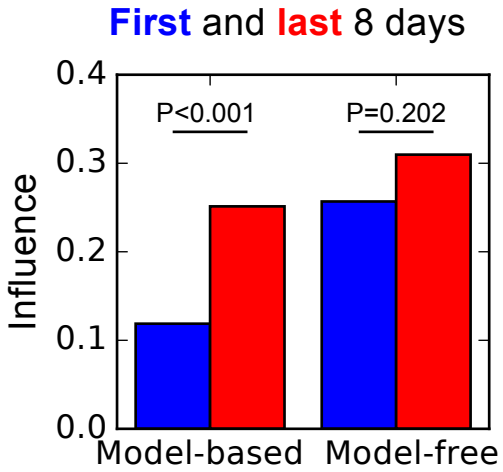


Figure 41: Reinforcement learning during the first and last eight days of the 52-day study. Model-based influence increased across sessions. P-values from within-subject permutation-testing.

In summary, performance and model-based learning increased across sessions (Table 8). This was measured by an increase in block end probability and loading on the *correct* predictor in the

logistic regression, meaning more correct choices, together with an increase in model-based influence on reinforcement learning measured by the RL model.

	Start of study → End of study
Trial activity	↑*** trial rate increase
Reaction time	↓*** after common transitions
Reversal Analysis	↑* block end probability ↓ ($P < 0.1$ for τ after reward probability reversals)
Logistic regression predictors	↑* correct, ↑*** choice, ↑** outcome, ↑* transition
RL model	↑*** model-based influence

Table 8: Summary of significant across session changes. *P<0.05, **P≤0.01, ***P≤0.001 by within-subject permutation-testing.

These changes in performance across sessions could potentially influence analyses of experiments. This will be revisited in section 4.3.3 where the effects of injection stress are analyzed.

4.3.2. Behavioural changes *within* sessions

To investigate possible *within* sessions behavioural changes, the 150 min sessions during days without injections were divided into the first 30 min, the mid 60 min and the last 60 min. As described in the previous chapter, subjects initially performed trials at a high rate (9 trials/min), which decreased to a steady state of approximately 3 trials / min after 20 minutes (Figure 42 left). During the session there is an initial high trial rate of 4.8±0.1 trials per min (±SEM), followed by a low trial rate of 2.9±0.1 followed by a minor increase to 3.1±0.1 (within-subject and cycle permutation-testing: first vs mid P<0.001, first vs last P<0.001, mid vs last P=0.001).

The 2nd step reaction times after common and rare transitions were different during the session (RM-ANOVA: F(1,7)=55.81, P<0.001), and both became shorter during the session (F(2,6)=25.21, P=0.001), but the difference in reaction time between the transition types decreased during the session demonstrated by an interaction of time period and transition type (RM-ANOVA: F(2,6) =

7.89, P = 0.021, Figure 43 right). Indeed, reaction time after rare transitions decreased after the first 30 min (Bonferroni multiple comparison: t(28)=8.15, P<0.001) and after 90 min reaction time after both common (t(28)=3.53, P=0.009) and rare transitions (t(28)=9.99, P<0.001).

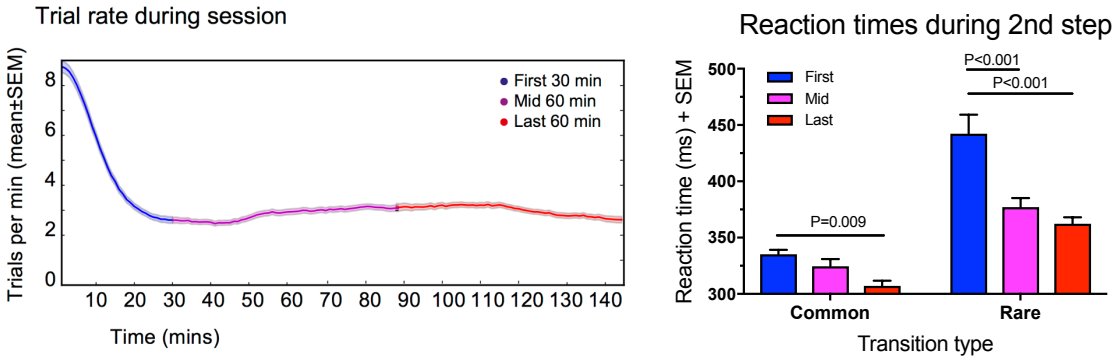


Figure 43: Mean(±SEM) trial rate and reaction times during the first 30 min, mid 60 min, and last 60 min of the sessions. Left) Mean (±SEM) cumulative trials during the first, mid and last period were 145±4, 171±5, 183±5, respectively. Right) Reaction times during the 2nd step depending on transition type.

We did a reversal analysis to identify changes in performance within sessions: τ after reward probability reversals decreased during the session (17.7→16.7→14.5 trials), but τ after transition probability reversals fluctuated during the session (13→12.4→15.5 trials), however, these differences were not significant by within-subject and permutation-testing. The mean block end probability significantly decreased during the sessions: 0.69→0.64→0.61 (within-subject permutation-testing: First vs Last: *P*<0.001, First vs Mid: *P*<0.001, Mid vs Last: *P*=0.013), indicating the subjects made a decreasing proportion of optimal choices at block end over the course of the session (Figure 44).

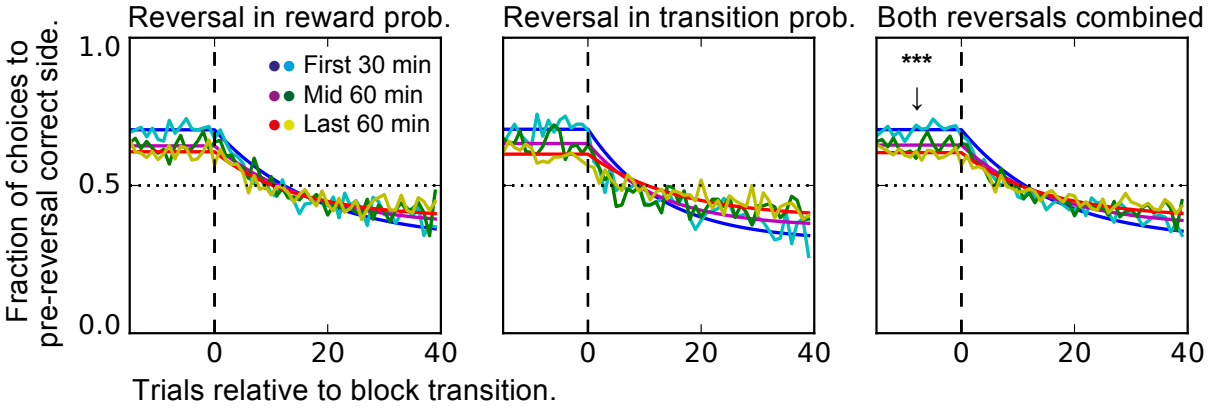


Figure 44: Reversal analysis within sessions. First 30 min → mid 60 min → last 60 min; τ after reward probability reversals: 17.7→16.7→14.5 trials, τ after transition probability reversals: 13→12.4→15.5 trials, τ after both reversal types: 16.2→15.4→14.8 trials, and block end probability: 0.69→0.64→0.61. ***Within-subject permutation-testing (First 30 min vs Last 60 min etc.: $P<0.001$).

Further investigation of choice behaviour by logistic regression revealed that loading on *correct*, *choice*, *outcome* and *transition* parameters declined during the sessions (Figure 45), which is in line with the decline in block end choice probability (Figure 44). This meant that after the initial 30 min the subjects had a decreased tendency to repeat choices following a correct choice (within-subject and cycle permutation-testing, correct: $P<0.001$) or a common transition (transition: $P=0.001$). After the initial 90 min, the subjects also had a decreased tendency to repeat rewarded choices (outcome: $P=0.015$), and a decreased tendency to repeat previous choices, independent of whether they were correct, common or rewarded (choice: $P<0.001$).

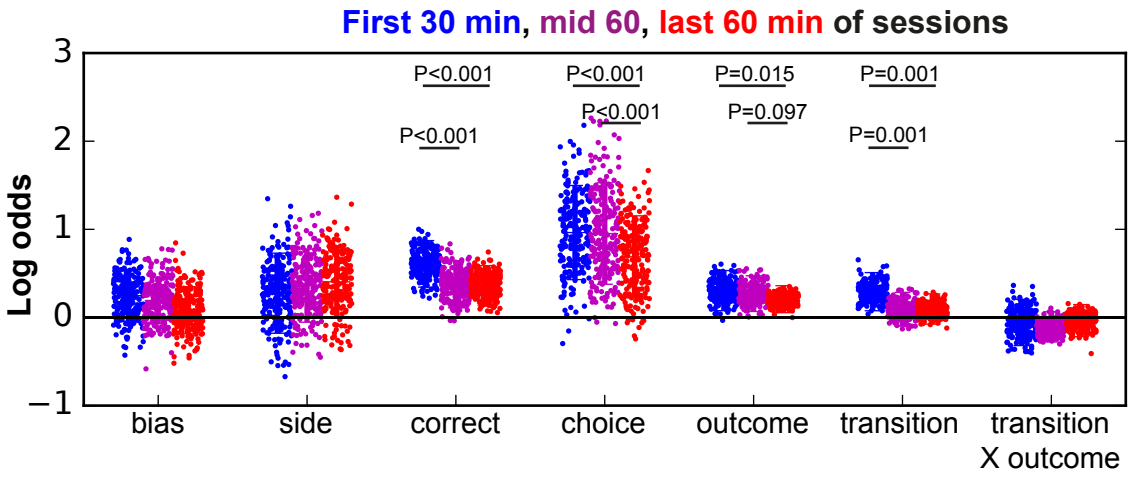


Figure 45: Logistic regression of choice behaviour during the first 30 min, mid 60 min, and last 60 min of sessions. Loading on the predictors *correct* and *transition* decreased after the first 30 min, and loading on *choice* after the first 90 min. Loading on *outcome* decreased when comparing the first and last part of sessions. P values derived using within-subject and cycles permutation-testing.

The RL model confirmed an initial large decrease in model-free influence after 30 min (within-subject and cycle permutation-testing: $P<0.001$) and a further although smaller decrease after 90

min (*P*=0.045). The model-free learning rate parameter (*alphaQ*) in the RL model declined significantly over the course of the session (First→Last: *P*=0.008). Model-based influence showed a minor decrease after the initial 30 min, but this was not statistically reliable (*P*=0.157).

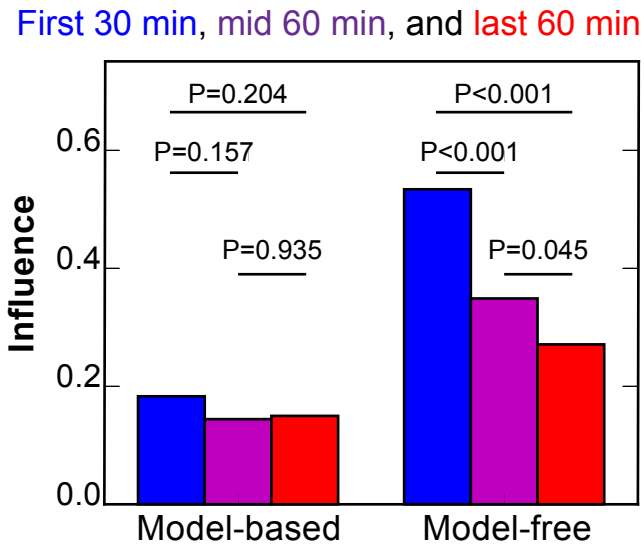


Figure 46: Reinforcement model of choice behaviour during the first 30 min, mid 60 min, last 60 min. The influence of model-free reinforcement learning declined during the session.

In conclusion, there was a significant within-session change: a decline in model-free reinforcement learning resulting in a decline in performance as seen by the drop in end block probability and a lower tendency to repeat correct choices (Table 9).

	First 30 min → Mid 60 min	**Mid 60 min → Last 60 min**
Trial activity	↓*** large trial rate decrease	↑*** minor trial rate increase
Reaction time	↓*** after rare transitions	↓** after rare and common trans.
Reversal analysis	↓*** block end probability	↓** block end probability
Logistic regression predictors	↓*** correct, ↓*** transition	↓*** choice, ↓** outcome[1]
RL model influence	↓*** Model-free influence	↓* Model-free influence

Table 9: Summary of significant within session changes. *P<0.05, **P≤0.01, ***P≤0.001 by within-subject and cycle permutation-testing. [1]For the last 60 min compared with first 30 min; Last 60 min compared with mid 60 min had *P*<0.1.

4.3.3. Stress study results

To study the effect of injection stress on reinforcement learning, performance on sessions after a vehicle injection, given either 15 or 90 minutes before testing, were compared with non-injection sessions. However, between-session changes introduced a problem: when permuting within-subjects of injection vs no injection, an injection session at the start of the study could be swapped with a non-injection session and at the end of the study, thus the effect of injection would be confounded by study date. To solve this, the within-subject permutations were limited to occur within short time frames (11 days) during which limited learning would occur, but including enough sessions to allow for the conditions to be permuted within the subject i.e. *within-subject and cycle* permutations. This essentially made each injection session paired with the time-wise adjunct five none injection sessions.

An injection 15 min before task initiation decreased the number of trials per session by 10% (316±8→282±13, within subject and cycle permutation-test: $P<0.001$), but an injection 90 min before did not affect trial rate ($P=0.651$) compared to no injection (Figure 47). The reaction times during

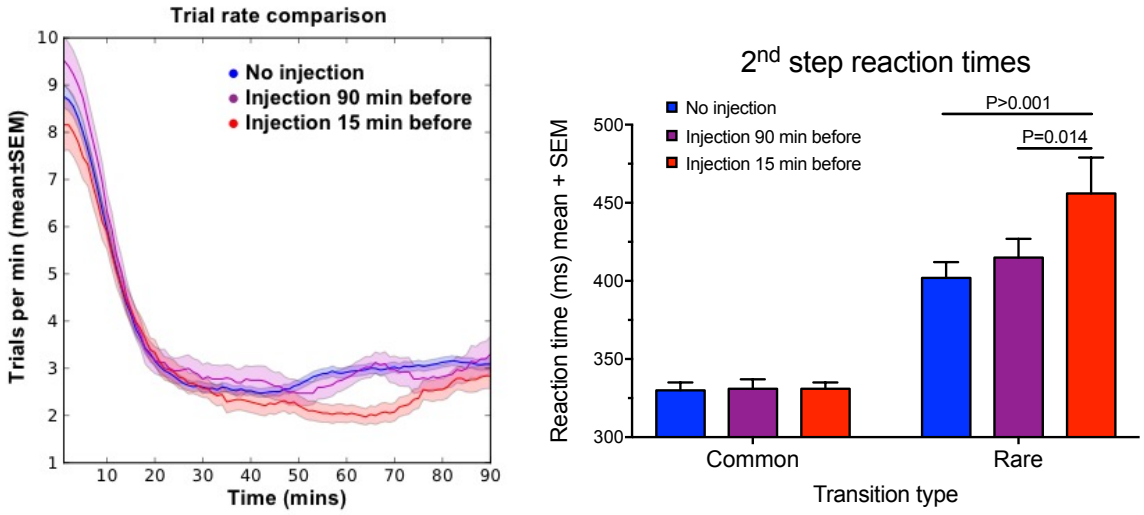
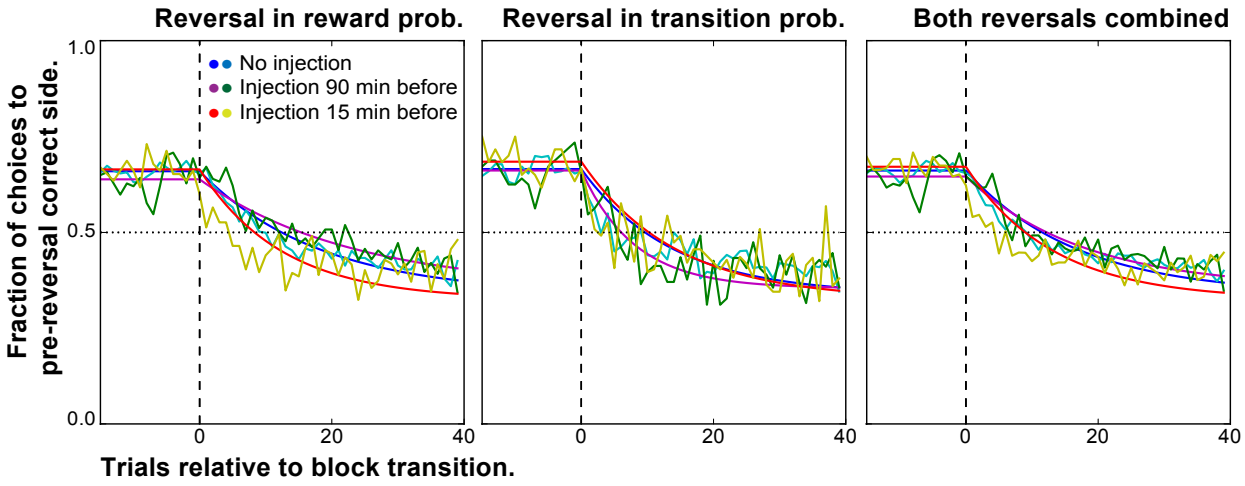


Figure 47: Trial rates and reaction times were affected by stress. (Left) Initial trial rate during the first minute (mean) and trials per session (mean± SEM) were: No injection: 8.8 trials/min and 316±8 trials, injection 90 min before: 9.5 trials/min and 324±18 trials, and injection 15 min before: 9.5 trials/min and 282±13 trials. (Right) Mean reaction times (±SEM) during the 2nd step. Reaction time following rare transitions was slower after an injection 15 min before.

the 2nd step following common transitions were not changed in either injection group, but an injection 15 min before increased reaction time following rare transitions compared to no injection (within subject and cycle permutation-test: $P<0.001$).

To study the effect of injection stress on performance, the sessions were compared by reversal analysis and logistic regression. τ after reward probability reversals was significantly shorter after an injection 15 min before (12.6 trials), compared with an injection 90 min before (23 trials; within-subject and cycle permutation-testing, $P=0.016$), while this approached significance when compared to no injection sessions (18.2 trials; $P=0.089$), suggesting increased performance when tracking changes in reward value after an injection 15 min before. All other differences in reversal τ were non-significant ($P>0.228$). Injections did not significantly affect the block end probability ($P>0.15$, Figure 48), nor were there any significant changes on predictor loading in the logistic regression ($P>0.15$, Figure 49).



	No injection	Inj. 90 min before	Inj. 15 min before
Average block end prob.:	0.66	0.65	0.68
Reward prob. reversal τ:	18.2	23.0	12.6
Transition prob. reversal τ:	13.7	8.1	14.2
Combined reversals τ:	16.8	17.7	13.2

Figure 48: Reversal analysis of sessions after no injections, an injection 90 min before, and an injection 15 min before. τ after a reward probability reversal was significantly shorter (from 23→12.6 trials) 15 min after injection compared to 90 min after injection ($P=0.016$, within subject and cycle permutations), and borderline significant compared to no injection (18.2→12.6; $P=0.089$).

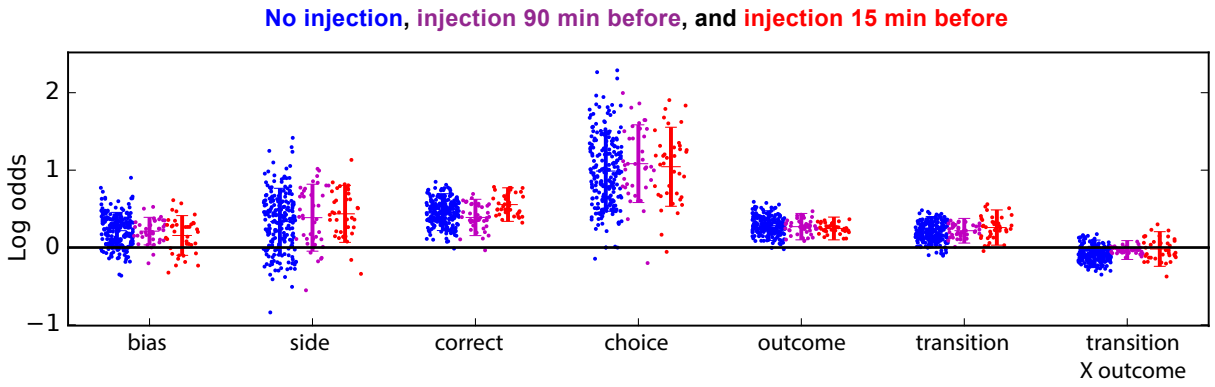


Figure 49: Logistic regression of sessions after no injections, an injection 90 min before, or an injection 15 min before. No significant changes on parameter loading by within-subject and cycle permutations was seen.

Human studies of stress have shown a shift from model-based to model-free reinforcement learning during stress (A. R. Otto et al. 2013b; Radenbach et al. 2015). In line with these studies, the RL model showed a reduction in model-based influence after an injection 15 min before compared with both an injection 90 min before (within-subject and cycle permutation testing: P=0.047) and no injection (P=0.040). Injection 90 min before had little effect on model-based influence compared with no injection (P=0.662). Moreover, an injection 15 min before increased model-free influence compared with an injection 90 min before (P=0.006) and no injection (P=0.005). An injection 90 min before increased model-free influence numerically, but not significantly, compared with no injection (P=0.121).

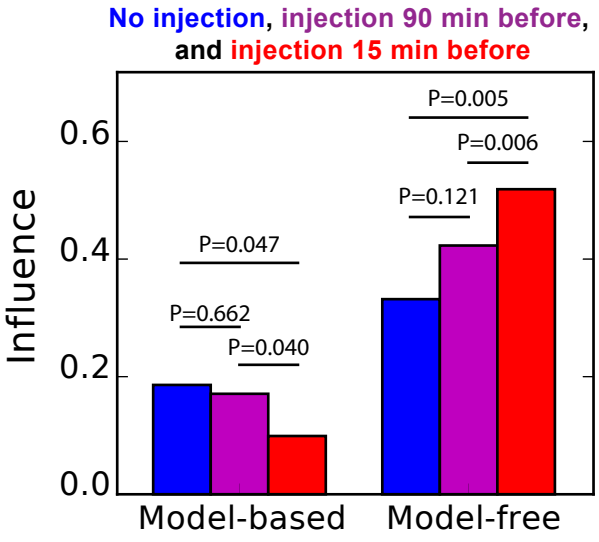


Figure 50: The RL model showed a change in strategy when exposed to injection stress. Stress by injection reduced model-based RL and increased model-free influence. Within-subject and cycle permutation testing.

In summary, stress from an injection 90 min before testing did not significantly affect any analysis parameters compared with no injection. In contrast, stress from the more recent injection 15 min before significantly changed several analysis parameters (Table 10).

	Injection 90 min before	**Injection 15 min before**
Trial activity	n.s.	↑*** trial rate increase
Reaction time	n.s.	↓*** slower on rare transitions only
Reversal Analysis	n.s.	↑P<0.1 τ after reward prob. reversals only[1]
Logistic regression predictors	n.s.	n.s.
RL model influence	n.s.	↓* Model-based influence ↑** Model-free influence

Table 10: Summary of significant effect of injections compared with no injection. [1]↓* τ after reward prob. reversals for injection 15 min before compared with 90 min before. *P<0.05, **P≤0.01, ***P≤0.001 by within-subject permutation-testing.

4.4. Discussion

This chapter studied the extent to which mice showed changes in the balance of decision-making strategies during reinforcement learning over different timescales and how performance was influenced following an acute stressor. Comparing performance at the start and end of a 52-day long study, there were significant increases in trial rate, performance and model-based influence across sessions during the study. Comparing the start, middle and end of session, there was a significant decline in trial rate, performance and model-free influence during the sessions.

Finally, comparing sessions 15 minutes after a vehicle injection, with those 90 minutes after a vehicle injection, or no injection, I found that acute injection stress decreased trial rate, increased performance on reward probability reversals and led to a shift towards model-free RL.

4.4.1. Changes across sessions

The results show that during the course of the study the mice increased their trial rate and performance, i.e. increased end of block choice probability and repeating the correct option, but did not statistically reliably increase the speed of reversals.

The logistic regression showed an increase in the tendency to repeat the correct option, but importantly also an increased tendency to repeat, dependent on both the previous rewards *and* the transition. Together, this suggests the performance increase was due to a more model-based behaviour. Furthermore, this conclusion was backed up by the RL-model. Across sessions the subjects became more model-based, with only an insignificant change in model-free influence on decision making.

The task is a complex instrumental task with sequential actions and with probabilistic rewards. The performance increase on the task was also apparent by the shorter τ after reversals, which meant subjects were faster to reach the threshold for block changes. This in turn meant more frequent probability changes and that the difficultly of the task increases with performance

increases. This would limit habitual responding and a purely model-free approach would lead to a limited performance increase, because while it is a fast computation it is an inefficient use of information as it disregards task structure and gives less accurate estimates. Indeed, because of its simplicity, model-free RL often exhibits slow learning in complex problems (Sutton and Barto 1998).

Therefore, truly increasing performance requires a shift from a model-free to a more model-based strategy as this takes into account the task structure and makes the best use of all the available evidence. This in turn gives more accurate estimates of which option is the best, but requires more experience with the task and its structure than a model-free approach would. Indeed, at the end of the study, when the subjects on average had experienced additional ~31.000 trials and ~300 reversals, the subjects had increased their model-based influence as apparent from the RL model, while there was no statistically reliable change in the model-free influence.

The changes in behaviour during this long study highlighted continuous changes in reinforcement learning. Even after setting rigorous performance criteria for the end of training, the subjects continue to improve and increase trial rate. This needs to be taken into account when studying other within-session effects. To control for across-session changes, the permutation testing of shorter-term effects in this chapter, i.e. within sessions and after injection stress, were limited to occur within cycles (i.e. time bins of 11 days) as also explained on page 53.

4.4.2. Changes within sessions

As well as observing changes across sessions, there were also changes in performance during individual sessions. First, the trial rate decreased markedly during the initial 20 min of the session. Moreover, there was a significant decline in performance, as measured by a lower tendency to repeat correct choices, both generally and also particularly after common transitions. This is in

line with a decrease in model-based influence. At the end of the session, there was a decreased tendency to repeat reward choices, and a further decrease in model-free influence.

One possibility was that the hunger experienced at the beginning of a testing session might have induced a mild stress, given that subjects had not had access to food for approximately the preceding ~21 hrs. However, the data did not indicate that this is stressful for the animals. As identified below and from human studies, stress impairs model-based learning: this was not observed during the initial 30 min compared to later in the session. In rodent studies, stress impairs outcome-revaluation (Braun and Hauber 2013; Dias-Ferreira et al. 2009), which is the opposite of what is observed.

Another hypothesis, that the later part of the session with lower trial activity allowed for more *computational* time for model-based learning was also not supported, since there was no change in model-based learning.

A decline in trial rate is expected in an appetitively-motivated task. After the first 100-150 rewards of sucrose the subjects become satiated and the initial reward value declines, and as a result, the motivation and the trials initiated to acquire the reward declined. The decision to make and take action, and which action, is a compound of model-free and model-based learning. The initial drop in trial rate and model-based learning is consistent with outcome-devaluation due to satiation and with classical instrumental studies of satiation (B. Balleine 1992; B. W. Balleine and Dickinson 1998; Colwill and Rescorla 1985).

But why does the influence of model-free learning also drop and continues to do so during the session? Model-free learning relies on RPEs and mesolimbic dopamine, which is shaped by motivation and specific properties of the reward to satisfy a current physiological need (Ahn and Phillips 1999; Bassareo and Di Chiara 1999; de Araujo et al. 2012; McCutcheon et al. 2012; Papageorgiou et al. 2016). Papageorgiou et al. observed a modulation of both food and sucrose-solution elicited dopamine in the nucleus accumbens of rats (measured by FCV) by the amount of

reinforcer consumed within and prior to the session (Papageorgiou et al. 2016). That mesolimbic dopamine and magnitude of the RPE declines within a session explains the continuous decline in model-free influence I observed. Together this supports the link between mesolimbic dopamine and physiological state on decision making and its role in prioritizing actions to maintain homeostasis.

Surprisingly, a re-examination of within-session changes in the water-reinforced cohorts did not find any significant within-session behavioural changes, apart from a decline in trial rate and a minor decrease in the tendency to repeat choices independent of outcome (Appendix 5.2.1). This suggests that our within-session changes may be specific to sucrose-reinforcement, and possibly due to its caloric value. However, in the water-reinforcement protocol, one other difference is that there is a dry food pellet present in the testing chamber to maintain motivation. This allows the subjects to disengage during the session to consume the food and reengage with a newly increased thirst. This in turn may mean that motivation and reward values do not show the same satiation in the water-reinforcement protocol as with the sucrose-reinforcement protocol.

4.4.3. Acute stress

We do not have physiological measurements of e.g. corticosterone or heart rate to quantify the stress levels of the mice during the task, but assume the mice are stressed during the 15-105 min period after injection based on several studies of mice that show that corticosterone or heart rate take up to 120 min to return to baseline after an injection (Drude et al. 2011; Meijer et al. 2006). From this we assume that when the mice are tested 90 min after injection, they are returning to their baseline no stress state, but may still have residual effect of the injection stress (Figure 35 left).

Acute stress from an injection 15 min before increased trial rate, but made the reaction times after rare transitions slower. This suggested that injection stress increased habitual responding

where the common action was more difficult during a rare transition. There was a decrease in model-based and an increase in model-free influences on learning (though without changes in performance), compared to no injection (Table 10). This discrepancy between the logistic regression and the RL model is due to a limitation in the logistic regression. A model-free influence increases the loading on the outcome predictor increases, while a decrease in the model-based influence decreases transition and outcome also, which in turn cancels out the effect of loading on outcome.

The same pattern of significantly reduced model-based and increased model-free on the RL model, but not change in the logistic regression, has also been observed in a study by Akam and colleges, who, in addition to the above, also argue that the limited change in transition loading may be due to compensating components, such as an increased model-free motor level perseverance that keeps transition from decreasing (T. Akam, personal communication, Sep. 2016). Thus, the logistic regression is not suited to observe the small change in the model-based component.

Though there were significant effects of stress on performance at 15 minutes, but not at 90 minutes, levels of model-based and model-free learning were numerically at intermediate levels between the two conditions, suggesting a time-dependent or stress-level effect (Figure 50). In short, this meant injection stress did not change task performance, but did mean a shift toward greater overall model-free control of decisions.

That acute stress induced a decrease in model-based influence and an increase in model-free is in line with human two-step studies (R. A. Otto et al. 2013d; Radenbach et al. 2015), and human and rodent studies of goal-directed and habitual control (Braun and Hauber 2013; L. Schwabe and Wolf 2009).

The results presented here are, however, the first rodent two-step data on the effects of acute

stress and allowed the subtle changes in model-based and model-free influence to be observed. The previous rodent study of acute stress on outcome-devaluation used a simple lever press task and found that multiple stressors, but not simple restraint or CORT injections, *shifted* instrumental responding from goal-directed to habitual (Braun and Hauber 2013). The study was limited by only being able to measure a shift in behaviour and thus unable to identify whether this was due to an impairment or enhancement of either goal-directed or habitual control, while we could observe both systems. The reason we observed an increase in model-free influence may be related to the debated arbitration between the two systems and could possibly be due to a disinhibition from the cortex on the striatum via corticobasal ganglia circuits (Akam T. 2016; Ito and Doya 2009) or enhanced striatal dopaminergic signalling during stress (Abercrombie et al. 1989; Wanat et al. 2013).

Future studies using multi-step decision making tasks could at last allow us to better understand the neurobiology underlying the use of different learning strategies in different conditions, such as those involving stress or mental disorders.

5. Separate and interactive effects of COMT genotype and stress on model-based/model-free learning

5.1. Introduction

Model-free learning relies on reward predictions and striatal dopamine circuits (Glimcher 2011), while model-based learning is considered PFC-dependent (R. A. Otto et al. 2013d). The enzyme catechol-O-methyltransferase catabolises prefrontal dopamine and a functional (Val^{148}Met) polymorphism leads to variation in the enzymatic activity, and consequentially also prefrontal dopamine tone and PFC function (Elizabeth M Tunbridge et al. 2006b). PFC function and model-based learning are impaired by acute stress. Therefore, variations in PFC dopamine and function by COMT could modulate reinforcement learning and/or the effects of acute stress. This chapter examines two questions:

- Does COMT genotype affect decision-making strategies?
- Does COMT-genotype modulate the effect of acute stress on decision-making strategies?

5.1.1. The COMT-Met mouse model

To investigate these questions, I used a mouse model of the human COMT Val^{158}Met polymorphism to investigate the modulation of prefrontal dopamine, and its interaction with stress, on reinforcement learning. The human Met allele was introduced into the native mouse COMT gene (Appendix 5.1) and COMT-Met mice are compared to wild-type littermates WT). WT rodent COMT activity shows activity similar to, or slightly higher, than human Val COMT (J. Chen et al. 2004; Lotta et al. 1995). Enzyme activity in COMT-Met mice is reduced by approximately 50% compared, and gene expression in the nucleus accumbens, dorsal striatum, and frontal cortex is not affected in COMT-Met mice (Barkus et al. 2016). Importantly, COMT-Met mice show normal performance on a range of basic behavioural measures, including locomotor activity, anxiety,

feeding behaviour and sucrose preference, and sensorimotor gating (Table 12, 2nd column), and little effect on standard memory and attentional performance at baseline. However, genotype differences emerge following administration of the COMT-inhibitor Tolcapone (Barkus et al. 2016). The COMT mice, therefore, provide a good basis for the exploration of reinforcement learning effects in the behavioural paradigms used here.

5.1.2. COMT and stress response

An overview of the different COMT models and studies is presented in Table 12 and Table 11).

Mouse	**[1]COMT-KO**	**[2]COMT-Met**		**[2]COMT-WT**	**[1]COMT-Val*tg***	**[3]COMT-OE**
Genotype	- /- & -/Leu	Met/Met		Leu/Leu	Val+++	Leu+++
E. Activity	0% & 50%	50%		100%	116%	130%
Rat				Leu/Leu		
Human		Met/Met	Met/Val	Val/Val		
Activity		60%	90%	100%		
Pharmacology		←Tolcapone / Amphetamine →				

Table 11: COMT-mouse models and their equivalent human genotype and effect of COMT-inhibition by Tolcapone. Wildtype mice and rats have a Leucine version with a similar enzymatic activity as the human Valine variant (Chen et al., 2004; Lotta et al., 1995). [1](F. Papaleo et al. 2008) [2] (Barkus et al. 2016) [3](Simpson et al. 2014).

In our lab, we have repeatedly seen differential behavioural effects of injection stress dependent on COMT status in rodents. Notably, lower COMT activity (either genetically in the COMT-Met mice or pharmacologically determined using systemic administration of tolcapone) appears to protect against the negative behavioural impact of injection stress. For instance, in the 5-choice serial reaction time task (5-CSRTT), COMT-Met mice maintained their performance to a greater extent than WT after saline injection, though this was only significant at trend level (Barkus et al. 2016). Similarly, on a progressive-ratio task, vehicle (20% cyclodextrin) injection reduced lever pressing compared to a pre-injection baseline day, an effect which was not observed when the COMT inhibitor Tolcapone was injected (C. Korn et al., unpublished observations).

Behavioural Task	[1]COMT-Met vs. COMT-WT	[2]COMT-KO vs. COMT-Valtg	[3]COMT-WT vs. COMT-OE	Rat WT vs. COMT-inhibiton
COMT activity	Met: ~50%, WT: 100%	KO: 0%, Valtg: 113%	WT: 100%, OE: 130%	-
Frontal Dopamine	-	-	OE: +80% peak, +85% DA release[7]	-
Locomotor activity	n.s.	n.s.	n.s.	-
Feeding behaviour				
Body weight	n.s.	n.s.	-	-
Sucrose preference test	n.s.	-	-	-
Anxiety related tasks				
Elevated plus maze	n.s.	Val-tg spent more time in the open arms than KO	n.s.	-
Light-dark box	n.s.	-	-	-
Novelty suppressed feeding task	n.s.	-	-	-
Anxiogenic open field	n.s.	n.s.	n.s.	-
Stress and pain				
Hot plate and tail flick tests	-	Decreased pain sensitivity in Val-tg and increased KO[5]	-	-
Stress-induced hyperthermia	-	Low in Val-tg; exacerbated in KO[5]	-	-
Sensorimotor gating				
Habituation to acoustic startle	n.s.	Val-tg showed lower acoustic startle reactivity than KO.	-	-
Prepulse inhibition	n.s.	n.s.	n.s.	-
Learning and memory				
Spatial novelty preference	Subtle increase in Met	-	-	COMT-inhibition improved performance[8]
Discrete trial spontaneous alteration	n.s.	-	-	-
Novel object recognition	n.s.	During the retention trial the Val-tg failed to spend more time exploring the novel object than KO.	-	-
Acquisition of appetitively-motivated reference Y-memory	n.s.	-	-	-
Morris water maze	n.s.	-	n.s.	-
T-maze test of working memory	-	Val-tg were impaired: KO improved[5]	-	-
5-choice serial reaction time task (5CSRTT)	Met outperformed WT at trend level on the long inter-trial interval stage. WT, but not Met performance improved by [4]COMT-inhibition	n.s.	OE showed more premature and perseverative responses, and slower acquisition, but once acquired were able to perform as well as WT.	-
Conditional associative learning paradigm	-	-	OE showed slower learning, but ultimately reached the same level of performance as WT.	-
Extra-dimensional set-shifting (EDS)	-	Val-tg mice required more time to solve the EDS task[6]	-	COMT-inhibition improved EDS[9]

Table 12: Summary of behavioural findings in rodent studies. n.s.: non-significant. [1](Barkus et al. 2016), [2](F. Papaleo et al. 2008) , [3](Simpson et al. 2014), [4]By Tolcapone, [5]Compared to WT. [6]Comapred to WT and KO pooled. [7]FCV of DA after stimulation of medial forebrain bundle under anaesthesia [8]Laatikainen 2012 [9]Tunbridge 2004. Due to sexually dimorphic effects of COMT (Desbonnet et al. 2012; Linda M. Laatikainen et al. 2013; F. Papaleo et al. 2012; Tammimäki and Männistö 2011) only results from male rodents are presented.

Nonetheless, COMT-Met mice do not show changes on ethological anxiety tasks (Table 12, 2nd column) (Barkus et al. 2016) (though see Papaleo et al. 2008 (Table 12, 3nd column) for contrasting results in complete COMT knock-out mice compared to COMT-overexpressing.

As outlined in the General Introduction, stress and dopamine signalling interact, and acute stress elevates cortical dopamine (A. F. Arnsten 2009a). From several human studies, there is an indication that the COMT Val158Met polymorphism impacts the cognitive and physiological responses to acute stress (Armbruster et al. 2011; Buckert et al. 2012; Desbonnet et al. 2012; Qin et al. 2009). Furthermore, the Inverted-U relation between PFC dopaminergic functioning and stress is suggested to be COMT-dependent (Buckert et al. 2012; Hernaus et al. 2013).

When these findings are considered in the light of the standard "Inverted-U" model of prefrontal dopamine function (Figure 51) this suggests that subjects with a higher baseline cortical dopamine availability (Met-homozygotes) will move to a sub-optimal dopamine level under stress, while subjects with lower baselines (Val-carriers) will be boosted to an optimal state. Given the proposed role of dopamine in working memory and model-based reinforcement learning, this might therefore influence the type of strategy the mice adopt.

Furthermore, dopamine in the nucleus accumbens core is elevated during acute stress (Inglis and Moghaddam 1999; Tidey and Miczek 1996). While dopamine in the ventral striatum, is not directly regulated by COMT, there is a range of evidence to suggest that COMT may nonetheless influence subcortical dopamine in certain conditions (E. M. Tunbridge et al. 2012), which may in turn influence model-free learning.

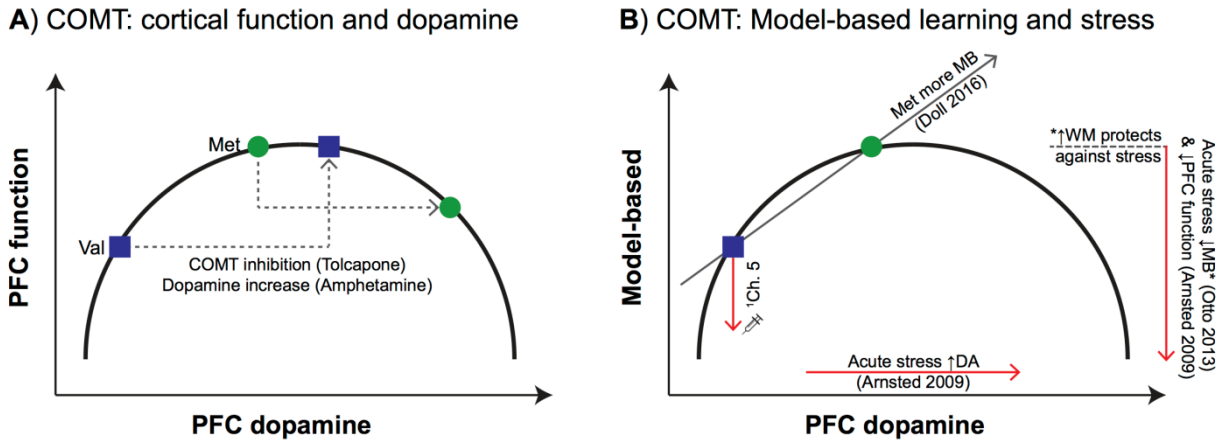


Figure 51: Cortical function, dopamine and COMT. A) The Inverted-U model of PFC function and dopamine. The Met allele with lower COMT activity has greater cortical dopamine compared with the Val allele. COMT inhibition or increase in dopamine is beneficial for the Val and detrimental to Met. B) Model-based learning (MB) and effects of acute stress (→). Acute stress increases PFC dopamine, but impairs PFC function (Amy F. T. Arnsten 2009b) and acute stress also impairs MB, *but this effect is limited by high working memory (WM) (R. A. Otto et al. 2013d). As suggested by the inverted-U and demonstrated on the two-step Met homozygotes are more model-based (Doll et al. 2016). In the previous chapter, ¹wildtype (Val homozygote equivalent) mice had decreased MB after acute (vehicle) injection stress (Chapter 5). Given the greater baseline level according to Doll, the Met homozygotes are expected to be less impaired on MB by injection stress.

5.1.3. Hypotheses

The two-step task involves continuously changing reward and transition probabilities and requires the ability to update and holds representations in the working memory is required. Humans carrying the high-activity Val allele tend to perform more poorly on tasks involving working memory (Egan et al. 2001; Mattay et al. 2003; E. M. Tunbridge et al. 2006a), and COMT-inhibition in humans has also been shown to improve working memory (Apud et al. 2007).

Cognitive flexibility is also required to adapt to the continuously changing probabilities. Extra-Dimensional Set-Shifting (EDS) is a measure of cognitive flexibility, which was poorer in COMT-Val*tg* compared to controls (WT+KO pooled)(F. Papaleo et al. 2012), while COMT-inhibition improved EDS in rats (E. M. Tunbridge et al. 2004). This suggests that COMT-Met mice may also have greater cognitive flexibility, which enables better adapting to the continuously changing probabilities of the two-step task and hence better performance on reversals at baseline.

Crucially, as suggested theoretically by the inverted-U and supported by the human two-step study, Met-homozygotes are more model-based (Doll et al. 2016). I therefore hypothesised that the COMT-Met mice, which are associated with putatively enhanced PFC function, cognitive flexibly, and computational capacity, will be more model-based than COMT-WT at baseline.

In the previous chapter, I found that wildtype mice had decreased model-based learning after acute injection stress similar to human two-step studies of acute stress (A. R. Otto et al. 2013a). We have also repeatedly seen COMT-genotype differential behavioural effects of injection stress in rodents, where COMT-Met mice appear protected from the injection stress in the progressive-ratio task and 5-CSRTT compared to WT. Based on this observation, combined with the greater baseline model-based level according to Doll, I hypothesised that COMT-Met mice would be less impaired by stress on the two-step task than the COMT-WT mice.

To test this, I conducted a similar stress study using the injection stress model as presented in the previous chapter. The 90 min analysis window from session start was retained to be able to compare COMT-WT mice with the C57BL/6 mice from the earlier study. As this was not part of a pharmacological study, the injections were saline rather than cyclodextrin, and were given 10 min, rather than 15 or 90 min before to capture the acute stress effects.

5.2. Methods

5.2.1. COMT training

Male, homozygous COMT-Met mice and their WT littermate controls (n=8 WT; n=8 COMT-Met) were divided into two groups of eight (n=4 COMT-Met and n=4 WTs per group, referred to as groups COMT.1 and COMT.2). The COMT mice were from 8 litters produced from heterozygous crosses (WT: 6 litters; COMT-Met: 5 litters). The COMT cohort had a mean age±SD of 20.3±1.2 weeks at the start of training and age was counterbalanced across COMT.1 and

COMT.2 (General Methods, page 31). Due to the scheduling of breeding and the study, the COMT cohort was 9 weeks older than the previously studied C57BL/6 cohorts.

The groups were run in parallel under the same conditions; COMT.1 was always tested in the early afternoon (12:15 pm – 3:30 pm) and COMT.2 afterwards (2:45 pm – 6 pm). They were weighed 30 min hours before testing and fed 30 min after testing. They were counterbalanced between genotypes in the testing boxes i.e. box 1 was used for a Met from COMT.1 and a WT from COMT.2.

Training on the two-step task started with reward delivery of 10 µl 20% sucrose solution, which was lowered to 6.5 µl on day 9, and to 4 µl on day 23. Session length started at 1.5 hrs and was extended to 2 hrs on day 3 and to 2.5 hrs on day 5. Transition and reward contrasts were lowered at day 15 and 21, respectively (Figure 52). Mice did not experience neutral reward blocks. The *persistent variables* feature was enabled throughout training and testing.

COMT.1 was inadvertently scruffed right before testing on day 26. To balance COMT.2 was therefore scruffed after testing the next day. To avoid the effects of stress confounding the baseline study, the data collection for the study was moved to start from day 28.

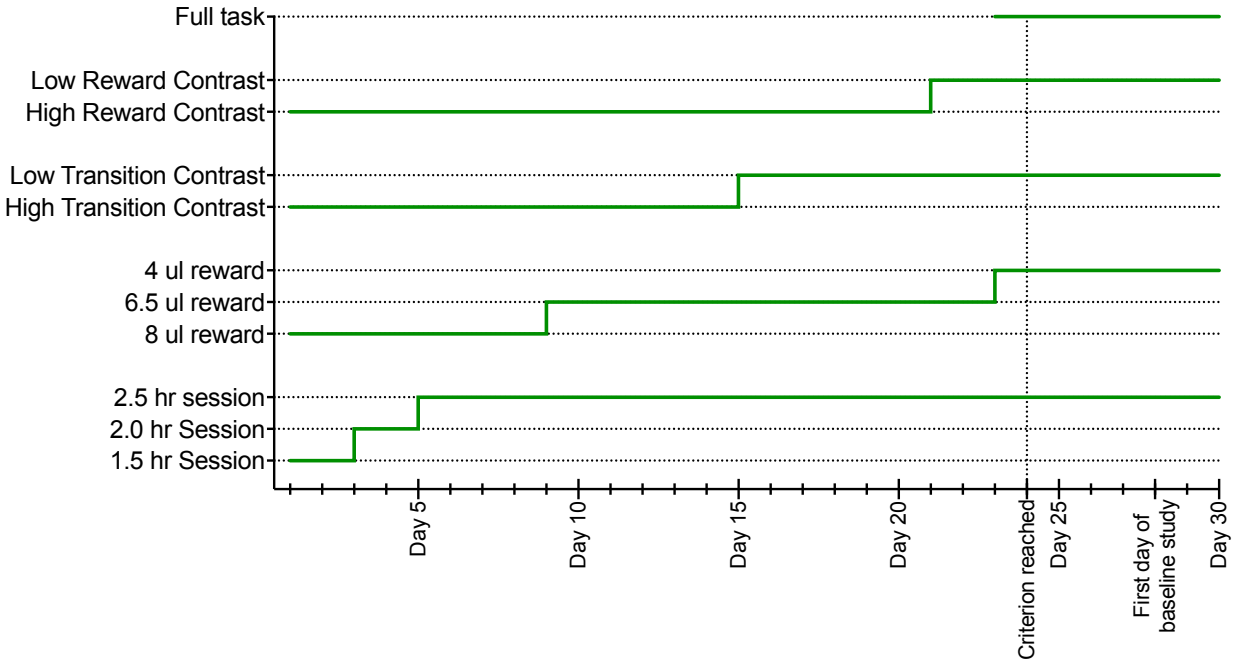


Figure 52: Summary of training of the COMT cohort.

5.2.2. Study of baseline effects of COMT genotype on performance

The baseline study was during day 28-48 with repeated testing 6 days a week (with no injections). Testing sessions were 2.5 hours, and the whole sessions were included in the analysis. The following sessions were excluded from analysis:

- Both subjects from box 6 (Met and WT) on day 37, due to a box malfunction.
- A WT subject on day 48, due to him escaping the box during testing.

Hence a total of 333 sessions (20.8 sessions/subject) were analysed.

5.2.3. The effect of injection stress on COMT-Met and COMT-WT mouse performance

The injection stress study was conducted between days 49-60, with alternating days of injection and non-injection days (Figure 53). After three days of no injection and injection, a day without testing was interleaved. Injections were alternated between the left and right side of each animal.

On injection stress days mice were weighed 15 min before testing, were scruffed and i.p. injected with saline (10 ml/kg) 10 min before testing. They were then moved to the testing room and placed in the testing boxes 5 min before testing. On non-injection days mice were weighed 30 min before testing in the holding room, and 5 min before testing were moved to the testing room and placed in the testing boxes.

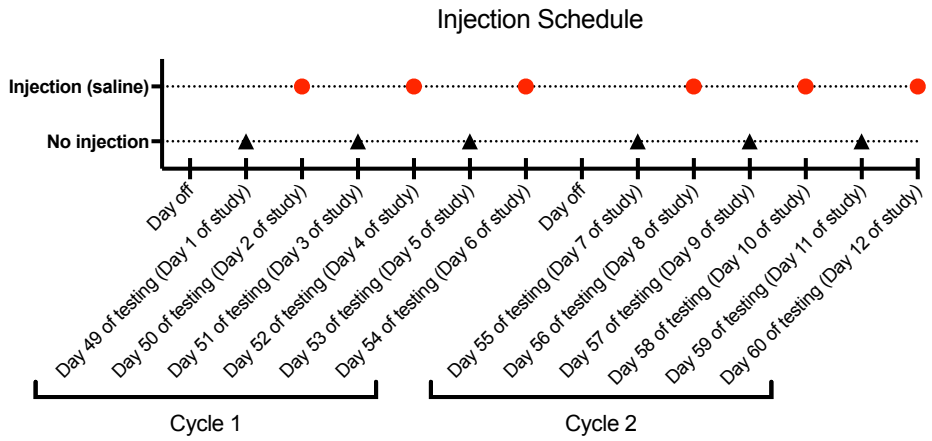


Figure 53: Injection schedule. Cycles were 6 days and preceded by a without testing.

The test start time was restricted more stringently than during the baseline period to limit the circadian effect of rising plasma corticosteroid during the afternoon (Appendix 6.2.1, (Malisch et al. 2008). COMT.1 was tested at 1:30 pm (range of start time: 12:15 pm – 2:30 pm) and COMT.2 at 4.15 pm (3:30 pm – 5:15 pm). Due to the extra preparation time involved, testing started on average 30 min later on days with injections than without for both COMT.1 and COMT.2. Testing sessions were 2.5 hours, but only the initial 90 min were included in the analysis. No sessions were excluded from analysis (Table 13).

Condition	No injection	Injection	Total
Sessions	96	96	192
Sessions/genotype	48	48	96
Sessions/subject	6	6	18

Table 13: Distribution of sessions COMT stress study

5.2.4. Statistics

Age, free feeding weight (FFW) and average percent of free feeding weight (FFW%) during training (until day 28) of the two genotypes were compared using unpaired two-tailed *t*-tests.

FFW% and the cumulative trials of the two genotypes at the start and end of (FFW% also during) the baseline and the injection stress studies were compared using an unpaired two-tailed *t*-test.

Changes in performance and reinforcement learning were analysed in the following way. Trial rate, 2nd step reaction times, reversal analysis, logistic regression and RL model described previously (General Methods) and statistically compared for an effect of genotype by cross-subject permutation-testing, and of injection stress by within-subject and cycle (6 days) permutation-testing.

Interactions of genotype and injection stress on performance was examined by three repeated measures ANOVAs (RM-ANOVA), with genotype as the between subject factor and injection the

within subject factor. The dependent variables were: 1) Block end probability, 2) Predictors from the single subject fitted logistic regressions, or 3) Influence of model-based and model-free RL from the single subject RL model fitting.

Note, the wild-type mice from the previous chapter, the SUCROSE 2 cohort, are not included in any analysis here. However, the SUCROSE 2 are here referred to as C57BL/6 mice, when paralleling the effects of acute stress on the COMT WT mice.

5.3. Results

5.3.1. Age, weight and training

We observed no differences between the COMT-Met and COMT-WT groups on age, FFW, or FFW% during day 1-28 (Table 14).

Genotype	**Size**	**Age (weeks)**		**FFW (*g*)**		**%FFW[1]**	
	n	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
COMT-WT	8	20.2	1.3	30.4	2.2	86%	2%
COMT-Met	8	20.4	1.1	30.9	2.6	86%	2%
Total	16	20.3	1.2	30.6	2.4	86%	2%
Unpaired t-test of genotype:		t(14)=0.33, *P*=0.75		t(14)=0.42, *P*=0.69		t(14)=0.00, *P*=1.00	

Genotype	**Cumulative trials at**					
	day 28		**day 48**		**day 60**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
COMT-WT	9,431	4,330	20,453	9,498	28,457	12,521
COMT-Met	7,968	2,797	16,154	6,022	22,779	7,548
Total	8,699	3,601	18,303	7,997	25,618	10,409
Unpaired *t*-test of genotype:	t(14)=0.80, *P*=0.44		t(14)=1.08, *P*=0.30		t(14)=0.33, *P*=0.29	

Table 14: Summary of baseline characteristics of COMT cohort. FFW: full fed weight. [1]During day 1-28. % of FFW during the baseline study (day28-48) was WT: 85.0%±2.4% and Met: 84.3%±1.8% (unpaired *t*-test: t(14)=0.66, P=0.52), and during the stress study (day 48-60) was WT: 82.4%±2.2% and Met: 82.4%±.2.6% (unpaired *t*-test: t(14)=0.00, P=1.0)

The COMT cohort was on the full task (i.e. low reward and transition probability and low reward size) on day 26. Both groups reached the performance criteria of a 3-day average of $\tau<25$ trials at day 24 to qualify for being ready for testing. But they were performing too few reversals

and trials per session to reach the other criteria (3-day average >4 reversals/session and 450 trials/session) to acquire a sufficient amount of data per session per analysis (Table 15).

However, the baseline study was designed to run over a minimum of 3 weeks. Thus, sufficient trials and reversals would be collected despite few trials and reversals per individual session. We, therefore, lowered the latter two criteria to ≈4 reversals/session and ≈400 trials/session to proceed with the next study.

	3-day average on day 24		
	Reversal τ	**Reversals/session**	**Trials/session**
Criteria - Original	*<25*	*>4*	*>450*
Criteria - Compromise	*<25*	*≈4*	*≈400*
COMT-WT	24.5	3.8	418
COMT-Met	24.4	4.2	373
Total	*24.5*	4.0	390

Table 15: 3-average at day 24 of τ for both reversal types, reversals/session and trials/session.

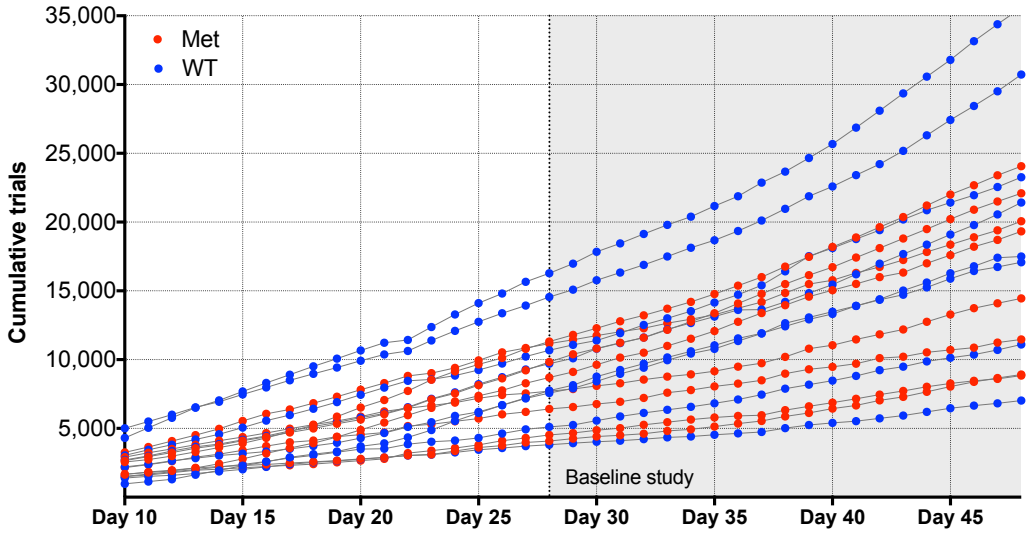


Figure 54: Cumulative trials by individual subjects during day 10-48. Light grey areas: day 28 – 48 of the baseline study. The two most active mice (i.e. highest cum. trials) were both WTs from the same litter.

5.3.2. Baseline effects of COMT genotype on performance

Figure 54 shows the accumulated amount of trials during the 21-day genotype study. The numeric difference in mean accumulated trials of the two genotypes was due to two WT littermates that had a high trial activity (Figure 54). However, the difference in accumulated trials

between the genotypes was not significant at the start (t(14)=0.80, *P*=0.44) or the end (t(14)=1.08, *P*=0.30) of the genotype study (Table 14). The WT and COMT subjects also did not differ in trials per session (cross subject permutation-testing: *P*=0.232) or reaction times (cross subject permutation-testing; Common: *P*=0.911, Rare: *P*=0.771, Difference: *P*=0.232, Figure 55). Since there was not effect on genotype on the trial rate or experience (i.e. accumulated trials), the subsequent analyses are not matched for accumulated trials.

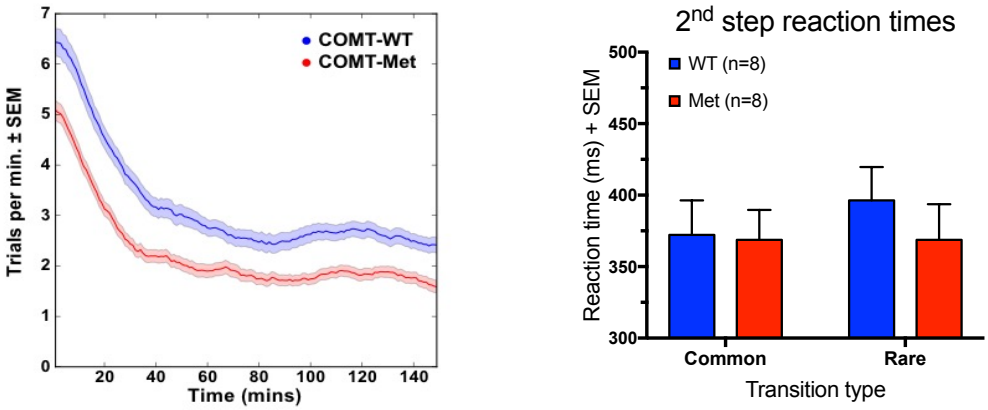


Figure 55: Trial rate and reaction time during the 2nd step. (Left) Mean±SEM trials/min of WT and Met (cross subject permutation-testing: *P*=0.232). (Right) Reaction time by transitions type (cross subject permutation-testing; Common: *P*=0.911, Rare: *P*=0.771, Difference: *P*=0.232).

The two genotypes were compared using the reversal analysis, logistic regression and RL model and cross subject permutation-testing described previously. There was no genotype difference on performance (τ) after probability reversals (cross subject permutation-testing: *P*=0.698), transition probability reversals (*P*=0.984), or both types combined (*P*=0.152), nor on the probability of selecting the correct option at the end of blocks (*P*=0.792, Figure 56).

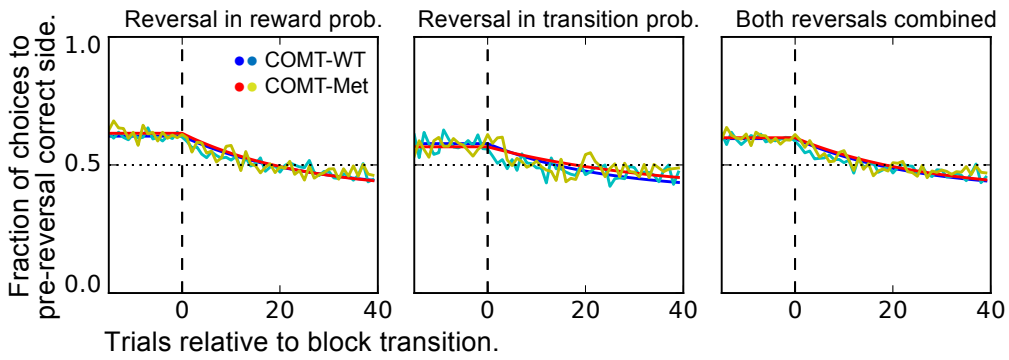


Figure 56: Performance on reversals was not dependent on COMT genotype.

When compared to the logistic regression, the Met group had a numerically lower tendency to repeat rewarded choices. However, this did not reach significance (cross subject permutation-testing*: outcome* P=0.075, other predictors P>0.38, Figure 57).

There was also no genotype difference on model-based or model-free reinforcement learning (cross-subject permutation-testing: both P≥0.347, Figure 58).

In summary, we observed no genotype differences at this point of training.

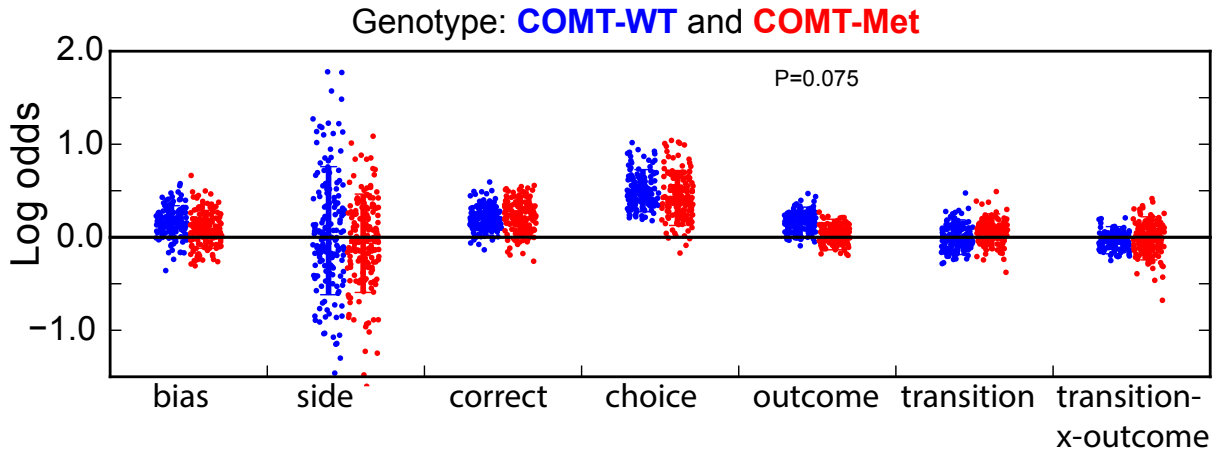


Figure 57: Logistic regression of choice behaviour was not dependent on COMT genotype. Despite a seemingly lower loading on *outcome* for the Met group, this was only significance at trend-level (P=0.075).

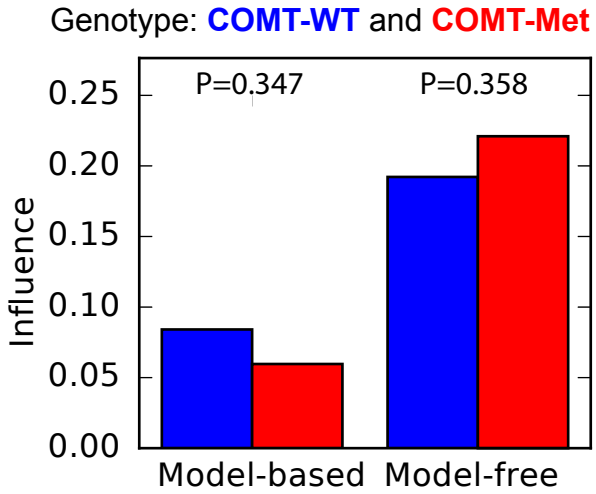


Figure 58: No significant difference in reinforcement learning dependent on COMT genotype. Cross-subject permutation-testing: model-based influence P=0.347 and model-free influence P=0.358.

5.3.3. The effect of injection stress on COMT mouse performance

Having not observed any genotype-depend differences in the previous section, we next investigated whether the COMT mice showed genotype-dependent responses to injection stress. Again, there was no genotype-dependent difference in FFW% during the study (P=1.0) or accumulated trials at the start or the end of this study (P's≥0.29, Table 14).

In the previous chapter we observed in wild-type mice (C57BL/6) an increased in trials per session, a decrease in model-based learning and an increase in model-free learning in the RL model. We therefore conducted the similar analysis to identify a difference between the COMT-Met and WT groups.

5.3.3.1. Trial rate and reaction times

An increased trials rate after injections was observed in the C57BL/6 mice in the previous chapter. A RM-ANOVA was conducted to investigate the within-subject effect of injection and the between-subjects effect of genotype on trial rate. There was a main effect of injection due to an increase in trials after injection ($F(1,14)=8.11$, P =0.0129) in the absence of a main effect of genotype ($F(1,14)=1.243$; P=0.2837), and there was an interaction between genotype and injection on trial rate ($F(1,14)=10.68$, P =0.0056). Post hoc tests using the Bonferroni correction revealed that WT mice increased their trial rate after injection (P=0.0014) similar to C57BL/6 mice, but Met were unchanged (P=1.0,

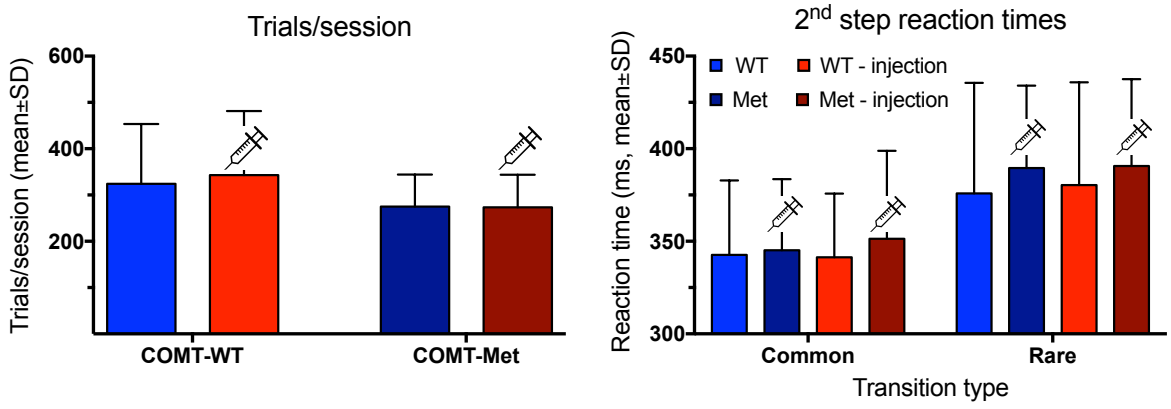


Figure 59).

A RM-ANOVA of the within-subject effect of injection and transition type and the between-subjects effect of genotype on reaction times showed only the expected main effect of a longer reaction time after rare transitions ($F(1,14)=46.215$, $P<0.001$) as in the previous chapter, while there were no other main or interactive effects ($F's<1.06$, $P's>0.32$,

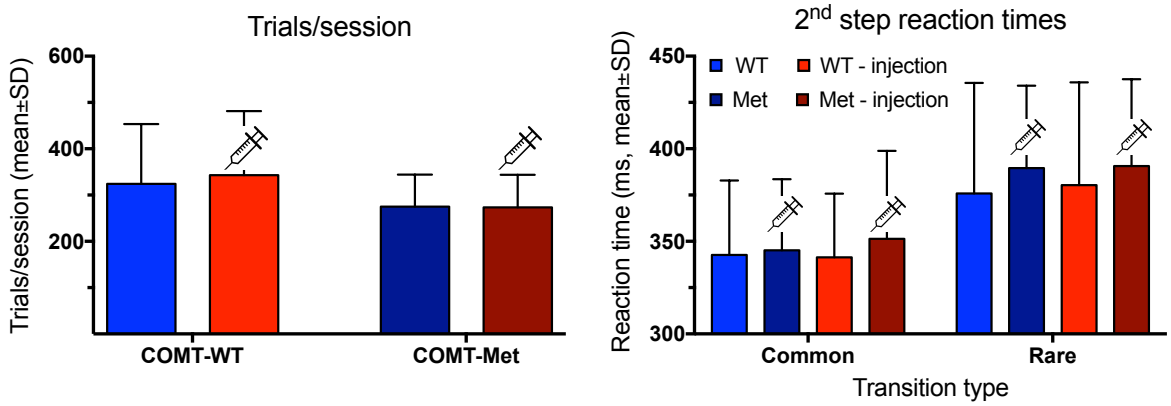


Figure 59).

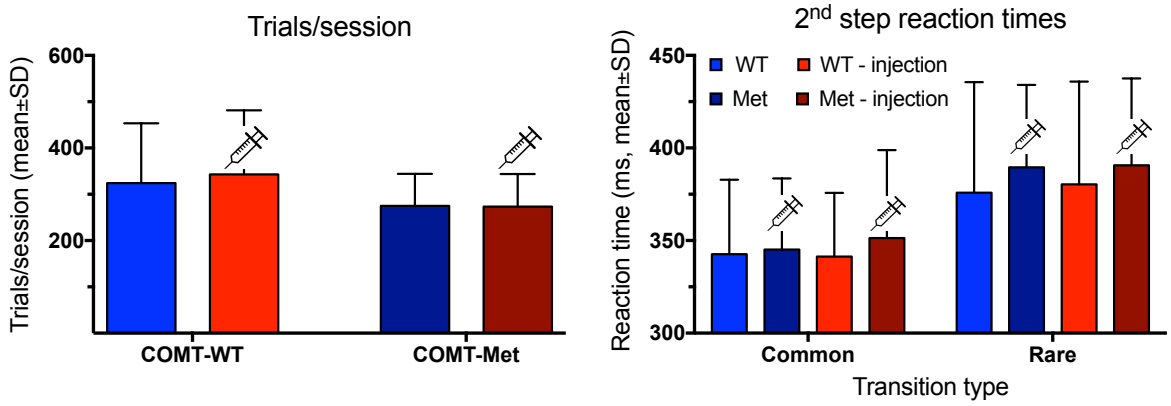


Figure 59: Trial rates and 2nd step reaction times by genotype and injection. Errors bare are SD of the mean across subjects of their mean across sessions.

5.3.3.2. Reversal analysis

There was no genotype effect on block end probability or τ after reward or transition probability reversals on days without injection (cross-subject permutation-testing: all $P>0.1$) as in the baseline genotype study.

Block end probability

The block end probability is a simple computation of the probability of correct choices during the 15 trials at the end of a block and could, therefore, be computed per subject. A RM-ANOVA was conducted to investigate the within-subject effect of injection and the between-subjects effect of genotype on the block end probability. There was a main effect of injection ($F(1,14)=10.15$, $P=0.0066$) due to an increase in mean block end probability after injection, and a main effect of genotype ($F(1,14)=4.808$; $P=0.0457$), due to an overall lower average performance at the block end in COMT-Met mice compared to the WTs. However, there was no interaction between genotype and injection on block end probability ($F(1,14)=0.9133$, $P=0.3555$).

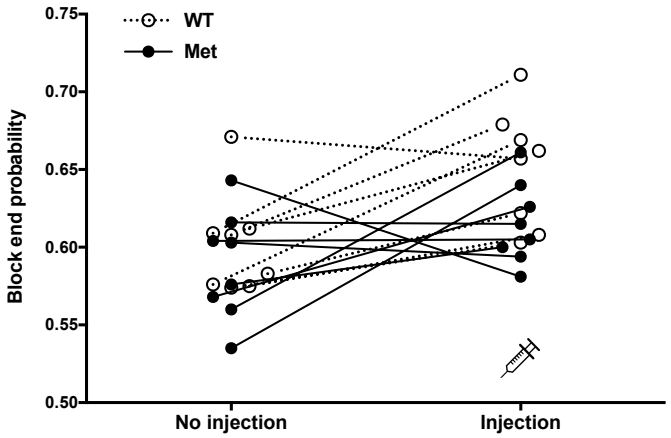


Figure 60: Block end probability increased after stress. Individual subject mean block end probabilities. WT had a higher block end probability after injection (0.63→0.68), than Met (0.61→0.63), the average increase for WT and Met was (0.60→0.64). There was no stress-genotype interaction ($P=0.355$).

Reversals

There were too few transition reversals for each subject to allow for single subject fitting of the τs meaning that a RM-ANOVA to compare the effect of injection and genotype on transition reversals was therefore not viable. However, there were enough per subject to give reliable fitting on the reward probability reversals (which happen twice as often as transition reversals). A RM-ANOVA of the within-subject effect of injection and the between-subjects effect of genotype on τ after reward probability reversals showed no main effect of injection ($F(1,14)= 0.989$, $P=0.34$) or

genotype (F(1,14)=0.215; *P*=0.65), and no interaction between genotype and injection (F(1,14)=1.87, *P* =0.19).

In the previous chapter a shorter τ after reward probability reversals was observed after injection, however this was only trend level significant (P<0.1). However, here, when the WT are analysed as whole similar to the previous chapter, there is no effect of injection stress (within-subject permutation-testing: P=0.874, Figure 61).

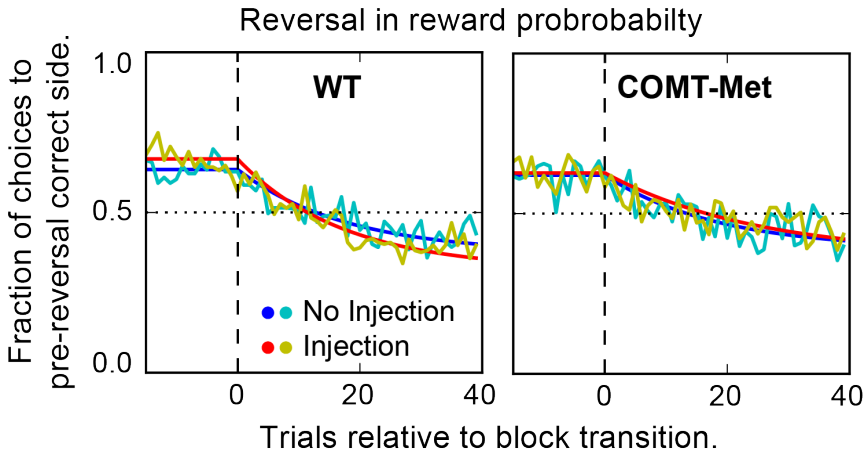


Figure 61 :Reward probability reversals after no injection (blue, cyan) and after injection (red, yellow). For τ was unchanged (17.9→17.3 trials, within-subject permutation-testing: *P*=0.874), for COMT-Met (18.5→25.0, *P*=0.299)

5.3.3.3. Logistic regression

There were no genotype differences in the logistic regression parameters on days without injection during this study (cross-subject permutation-testing: all predictors $P \geq 0.1$, Figure 62 blue plots), confirming the findings of the baseline genotype study.

The possibility of interactive effects of genotype and stress on the loading of the predictors in the logistic regression was tested using an RM-ANOVA on single-subject fitted regressions. The ANOVA compared the within-subject effect of injection and between-subjects effect of genotype on the predictors. There was an interaction between genotype and stress on the predictor *transition* loading (F(1,14)=4.648, *P*=0.049), and a trend-level main effect of genotype due to a reduced

loading in WT after injection compared with an increased loading in Met (F(1,14)=4.363, P=0.055), while a main effect of injection stress was absent (F(1,14)=0.76, P=0.398. While there were no significant effects or interactions on the *choice* parameter, the interaction between genotype and injection on approached significance (F(1,14)=3.482, *P*=0.083; main effects of injection: F(1,14)=2.082, P=0.171; choice: F(1,14)=0.565, P=0.465). There were no other interactive or main effects for the logistic regression predictors (F's<2.082, *P*'s>0.17).

To further examine the interactive effects of genotype and stress, the logistic regression was fitted separately for each genotype and condition (with/without injection), and the effect of injection stress was compared separately for WT and MET by within-subject permutation-testing (Figure 62). This confirmed that the Met group had a greater tendency to repeat choices following common transitions (*transition*, *P*=0.010); in contrast, the WT group had no reliable change (*P*=0.147). Such an effect was similarly observed in C57BL/6 mice in the previous chapter i.e. no change on *transition* loading. WTs had an increased tendency to repeat choices after injection, independent of outcome or transitions (*choice*, within-subject, *P*=0.029. By contrast, the Met group did not show a change in tendency to repeat choices independent of outcome or transitions (*P*=0.988). There were no changes in other predictors in either genotype (*P*'s>0.1).

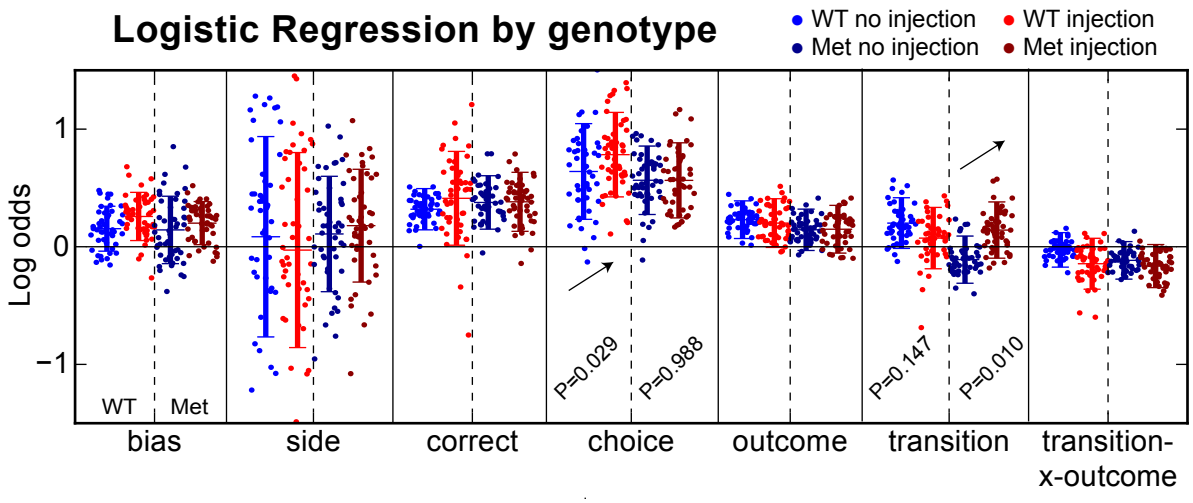


Figure 62: Logistic regression of choice behaviour after no injection and injection divided by COMT genotype (WT left; light blue and light red, Met right; dark blue and dark red). No genotype difference in the logistic regression on days without injection (cross-subject permutation-testing: all predictors *P*≥0.1). In

response to the injection, WT had an increased tendency to repeat choices independent of outcome or transitions (*choice*, within-subject permutation-testing: *P*=0.029), which the Met group did not have (*P*=0.988). The Met group had a significantly lower tendency to repeat choices following common transitions (*transition*, *P*=0.010), while the WT group had a non-significant decrease (*P*=0.147).

In short, the WT group repeated more choices independent of outcome or transition, while the Met group repeated choices after a common transition after injection compared to no injection. Moreover, there was a significant genotype-injection interaction on repeating common transitions, but not on other predictors. The divergent choice behaviour suggested a possible subtle genotype-modulated response to injection stress in reinforcement learning.

5.3.3.4. RL model

There was no reliable genotype difference in the RL on days without injection during this study (cross-subject permutation-testing: model-free infl. *P*=0.329, and model-based infl. *P*=0.078, Table 16), confirming the findings of the baseline study. Therefore, to study the possible genotype dependent stress effects on reinforcement learning, RL modelling was conducted separately for each genotype, and sessions after an injection or no injection were compared by within-subject permutation-testing (

Figure 63).

We observed an increase in model-free learning (*P*=0.018) in WTs after injection as well as a trend for the opposite effect on model-based learning (*P*=0.068). This pattern was comparable to that observed following injection stress in C57BL/6 mice in the previous chapter. In contrast, model-based (*P*=0.918) and model-free learning (*P*=0.766) were unchanged after injection versus non-injection in the Met mice.

This suggested that COMT-genotype influences patterns of RL in response to acute stress. However, this type of analysis cannot say whether there is a significantly greater influence of stress in the WT than the Met mice. Therefore, to investigate this we ran a companion analysis

using extracted parameters for each individual mouse and subjecting these to a RM-ANOVA (within subject effect of injection and between subject effect of genotype on the influence of model-based and model-free RL). However, none of the main effects or interactions were significant (F's< 2.6, P's>0.13).

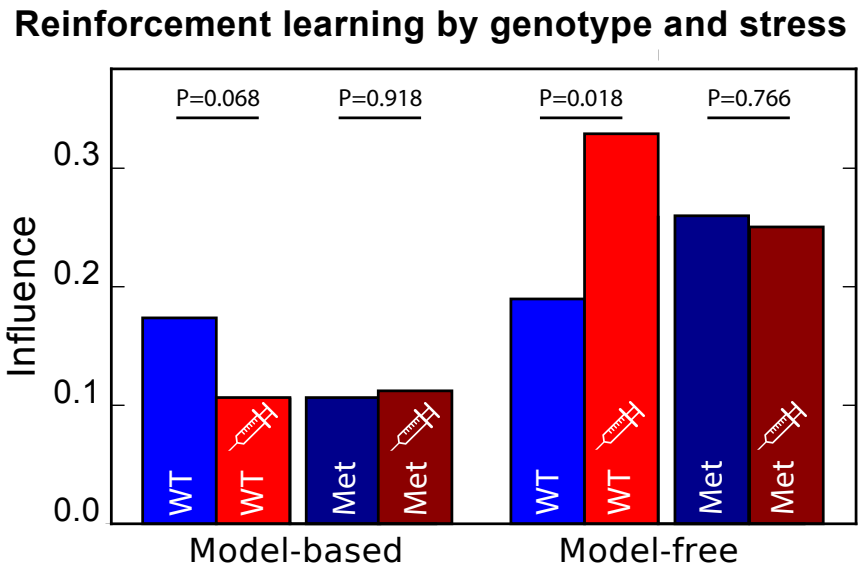


Figure 63: RL model of choice behaviour following no injection (blue) and injection (red) divided by COMT genotype (WT left; light blue and light red, COMT-Met right; dark blue and dark red).

		WT		**COMT-MET**		*Cross-subject*
		MB	MF	MB	MF	*P=*
No-injection	MB	0.168		0.109		*0.078*
	MF		0.105		0.261	*0.329*
Injection	MB	0.190		0.115		*0.791*
	MF		0.330		0.278	*0.340*
Within-subject	*P=*	*0.068*	*0.018*	*0.918*	*0.766*	

Table 16: Summary of RL model comparisons. Permutation-testing was within-subjects to compare the effect of injection stress and cross-subject to compare the effect of genotype.

5.3.4. Brief summary of findings

	Effect of genotype during baseline study[1]	**Effects of injection by genotype[1]**		**Injection-genotype interaction[3]**
		WT	**COMT-Met**	
Trial activity	n.s.	↑*** increased trial rate	n.s.	** trial rate

Reaction time	n.s.	↑*** slower after rare transitions[2]		n.s.
Reversal Analysis	n.s.	↑** block end probability increase[2]		n.s.
Logistic regression	n.s.	↑* choice	↑** transition	* transition P<0.1 choice
RL model influence	n.s.	↓P<0.1 model-based ↑* model-free	n.s. n.s.	n.s.

Table 17: Summary of findings of the effect of stress by COMT genotype. *P<0.05, **P≤0.01, ***P≤0.001 [1]By within-subject permutation-testing. [2]Main effect of injection from RM-ANOVA. [3]RM-ANOVA.

5.4. Discussion

5.4.1. Genotype study

In this study, I investigated how COMT-genotype affected performance and reinforcement learning. I hypothesised that Met homozygotes, which are predicted to have increased PFC DA, would be more model-based than WT homozygotes. However, the results show no genotype differences in trial rate, reaction time, reversal learning, correct responding, or influence of model-based and model-free reinforcement learning.

Although this is not what was originally hypothesised, the results appear largely consistent with other unpublished work in the lab investigating the effects of pharmacological COMT inhibition in wild type (C57BL/6) mice that also showed no change in the influence of model-based and model-free learning (C. Korn, personal communication, May, 2016).

Model-based learning

Both the human and rodent two-step study find no genotype effect on model-free learning (Doll et al. 2016). However, our results are in contrast to the findings of a Met gene-dose dependent increase in *model-based* learning in humans (Doll et al. 2016). They observed a linear effect of Met allele number on model-based decision making in a logistic regression (at P=0.01), but no effect on effect on model-free decision making, and an increased model-based choice weight in their RL model (at P=0.03). There are several differences that may explain this contrast summarised in Table 18.

First, it is possible that we may be overlooking a possible genotype difference due to a limited amount of subjects in our study. Further work with larger cohort numbers will be helpful to determine this. Second, Doll et al. included young women in the study and the results of may be influenced by possible sexual dimorphic effects of COMT as observed in both human and rodent

studies: COMT expression is inhibited and cortical dopamine is enhanced by oestrogens, and COMT activity appears to be lower in females than in males (J. Chen et al. 2004; Gogos et al. 1998; Harrison and Tunbridge 2008; Linda M. Laatikainen et al. 2013). Notably, working memory, an essential part of model-based learning, has shown sexually dimorphic effects of COMT genotype. A recent study found that COMT modulated WM (*n*-back task) in men, but not in premenopausal women, and that the sex-dichotomous effects disappeared after menopause (Francesco Papaleo et al. 2015). In premenopausal women, the direction of effects of endogenous oestradiol on working memory (also *n*-back task) was dependent on COMT genotype (Jacobs and D'Esposito 2011). Third, the rodent task may not be demanding enough to observe genotype differences that are dependent on task demand i.e. if the task is too simple the greater ability of one group is not revealed, similarly a task too difficult for both groups may not reveal a differential ability. Bucket et al. 2012 observed such an effect of task demand, where there COMT-genotype differences in humans performing the 2-back test under psychosocial stress, but Met and Val-homozygotes were equal on the easy 1-back and the more difficult 3-backversion (Buckert et al. 2012).

Finally, the rodent two-step task is a reduced version of the human task with one rather than two choices and two rather than four outcomes. In the human version the value probabilities change each trial which requires the subject to update values estimates continuously, whereas the rodent version has value probabilities in blocks allowing the subjects to aggregate value information over several trials (General Methods). Consequently, the rodent task may not properly reflect the cognitive processes in the human version. Moreover, humans can do the task without training in a model-based and model-free way (participants are usually given 30 trials to familiarise with the task), whereas mice take 8-10.000 trials to properly learn the task. In essence, the model-based capacity of mice may just not be directly comparable with humans.

While we, unlike Doll, do not find genotype differences in model-based learning at baseline, they may emerge under more other conditions, such as stress.

	Doll et al. [1] *Human two-step*	**Our study** *Rodent two-step*[2]	**Caveat**
Outcome changes	Each trial	In blocks[3]	Difference in structure and complexity of task and models
Logistic regression	For human two-step	For rodent two-step[2]	
RL model	7 parameter model	14 parameter model[2]	
Subjects	Male and female (57%)	Male	Sexual dimorphic COMT effects[4]
n (M/M; M/V; V/V)	171 (56, 80, 33)	16 (8, - , 8)	Difference in data size
Trials per subjects	300 trials	~1.800 trials	

Table 18: Difference in the human and rodent COMT two-step studies. [1](Doll et al. 2016), [2](Thomas Akam et al. 2015) [3]See General Methods. [4]Sexual dimorphic COMT effects is observed in several human and rodent studies (see main text).

To perform adaptively given the continuously changing reward and transition probabilities in the two-step task, the ability to update and holds representations in the working memory is beneficial. Humans carrying the high-activity Val allele tend to perform more poorly on tasks involving working memory (Egan et al. 2001; Mattay et al. 2003; E. M. Tunbridge et al. 2006a). Similarly, COMT inhibition in rats and improves performance on hippocampal-dependent memory tasks (e.g. spatial novelty preference task)(L. M. Laatikainen 2010; E. M. Tunbridge et al. 2004), which some groups argue contains a working memory component, while mice overexpressing the human Val variant (COMT-Val-tg) had impaired working memory (T-maze learning) compared to COMT-knockout mice (F. Papaleo et al. 2008).

Yet, it is not clear if or how COMT influences working memory in rodents, but we see no difference in model-based learning in the Met and WT mice. Impaired working memory could in principle translate into impaired model-based learning. However, it is important to remember that the two-step task is not simply a test of working memory alone, as model-based learning relies on other neuronal components.

Model-free learning

The COMT-Met mice also showed no difference from WT mice in their sensitivity to reward, as measured in terms of the motivation to perform trials, the outcome predictor in the logistic regression, or a bias towards model-free influence in the RL model. This was in line with the

aforementioned studies of COMT-genotype in humans (Doll et al. 2016) and COMT–inhibition in C57BL/6 mice (C. Korn, personal communication).

Our results contrast with Lancaster and colleagues' four human studies (predominately women) that found Met homozygotes compared to Val carriers have increased reward learning measured as a bias towards repeating rewarded choices in a simple probabilistic reward learning task where the participant answered yes or no to an alternating onscreen stimulus (Corral-Frias et al. 2016). The role of COMT in reinforcement learning seems to depends on task requirement: In a task with fixed outcome contingencies (i.e. stable reward representations in WM), the Met allele was favourable in human participants (Frank et al. 2007), however, in a task with changing outcome contingencies (i.e. requiring continuous updating by striatal RPEs) the Val allele was favourable (Krugel et al. 2009).

Indeed, there are a number of studies showing an influence of COMT on reinforcement learning, but the nature of this effect in most studies is not yet clear. Therefore, while it is somewhat surprising to see no difference, it might be related to the type of tasks with regards to task demand as mentioned earlier and which neuronal circuits the task taxes.

5.4.2. Genotype – Stress interaction study

In this study, I investigated if COMT-Met mice were less impaired by acute stress on the two-step task than the WT mice. To test this, I conducted a similar stress study using the injection stress model as in the previous chapter.

The results show that, following injection stress, the WT mice increased their trial rate similar to the previously studied C57BL/6 mice (Chapter 4), but the Met mice did not, and there was a positive genotype-stress interaction. The WT mice also had an increased tendency to repeat choices independent of previous rewards or transitions, that in turn resulted in higher correct choices at the end of blocks, but no actual performance increase as measured by τ on reversals, suggesting a

shift towards automatic responding. There was an interaction on the predictor loading on transition and choice (although the latter should be treated with caution, given that P<0.1). Together, this suggests that after injection stress WT mice shifted to a habitual strategy of repeating choices irrespective of reinforcement and transition structure, while by contrast the Met mice became more sensitive to the task structure.

In the RL model the WT showed a significant increase in the weight of influence of model-free learning along with a numerical (non-significant) reduction in the influence of model-based learning in response to acute injection stress. This pattern was very similar to that observed in the 10 week younger C57BL/6 mice in the previous chapter after an injection 15 minutes before testing, and is in line with human two-step studies that also find a model-based impairment and a shift towards model-free RL in response to acute stressors (cold pressor stress test and TSST, (R. A. Otto et al. 2013d; Radenbach et al. 2015). These findings converge with earlier work showing that stress impairs devaluation sensitivity in both humans (L. Schwabe and Wolf 2009) and rats (Dias-Ferreira et al. 2009).

That the WT did not reach $P<0.05$ (but only $P=0.068$) on the effect of injection stress on model-based as their C57BL/6 counterparts ($P=0.046$) could be due to age difference or that COMT mice were are tested later in the day when their basal stress hormone, CORT, levels are higher and thus less effected by injection stress. However, differences in power are more likely: this study had only 1/5 of the total amount of sessions of the previous study[1] (1.25 more injection sessions per subject, but 15.8 fewer none injection sessions per subject than in the previous chapter).

WT, similar to the C57BL/6, showed a discrepancy between the logistic regression showing no changes on predictor loading on outcome, transition, while the RL model showed a model-based decrease and model-free increase. Briefly, this is due to a limitation in the logistic regression,

[1] The frugality was due to a planned subsequent study in the COMT-mice not presented here.

where an increase model-free increases on the outcome and decreases transition, which is cancelled out by the coinciding opposing effect on a model-based decrease.

Interestingly, the COMT-Met mice were much less affected by injection stress and did not show significant changes in either model-based or model-free learning. However, when single subjects fitted values from the RL model were exported into an ANOVA, there were neither genotype-injection stress interaction, nor any main effects of injection or genotype, and only a trend level effect of injection on model-based learning, meaning that the differences between the groups need to be treated with caution. However, it is important to note that there is a considerable loss of power when fitting the RL model on 6 sessions per subject per condition rather than the 48 sessions of the group per condition and taking full advantage of the multilevel modelling approach and the statistical approach of permutation the whole dataset (256 sessions). Therefore, the lack of any significant interactions should not definitively be taken as evidence for a lack of differences between the groups.

Permutation testing gives a better resolution of the differences in the dataset, but an exact test calculating all possible permutations for interaction cannot be calculated (Manly 2007b). However, there are good approximate tests using permutation testing and Monto Carlo process (Anderson 2001; Manly 2007a). This requires custom-written code for each test (i.e. the reversal analysis, logistic regression, and RL model) to be run on parallel nodes on a computer cluster (i.e. a super computer), which, while an obvious next step, is unfortunately beyond the scope of this thesis.

A stress-genotype interaction was proposed by Buckert (2012) regarding the observation that Val homozygotes, compared to Met-homozygotes, performed better on the n-back (on 2-back level, not 1- or 3-back) test of WM after an acute psychosocial stressor (Buckert et al. 2012). The n-back test only tests working memory, which is important to model-based learning, but does not test aspects of reward learning or cognitive flexibly like the two-step task. Moreover, while both model-based and model-free RL are dopamine dependent (Daw et al. 2011; Doll et al. 2016),

increased dopamine by L-DOPA to participants (of unknown COMT genotype) increased model-based learning (Sharp et al. 2016). And acute stress observed here, in animals and in human studies show impairments in model-based learning (A. R. Otto et al. 2013b). It is simplistic to assume, that stress alone increased prefrontal dopamine and mediate the cognitive impairments.

The finding that COMT-Met mice, in contrast to WT mice, appear unaffected by stress is in opposition to Buckert's theory: If stress *uniformly* increases PFC dopamine, this should move WT subjects toward the peak of the hypothesised Inverted-U to a more optimal cortical DA level (Wunderlich et al. 2012), whilst Met subjects would be pushed to a supra-optimal level of PFC dopamine and be negatively affected by acute stress (Buckert et al. 2012). This is based on Buckert et al.'s observations that human Val-homozygotes performed better than Met on the *n*-back test (on 2-back level, not 1- or 3-back) after an acute psychosocial stressor (Buckert et al. 2012). This is not what I observe regarding model-based RL in the COMT mice.

The *n*-back test only tests WM, which *is* important for model-based learning. However, the reason may be that the two-step task is dopamine-dependent in a way previous WM tasks, and taxes not only WM networks, but involves reward learning circuits.

One speculation is that while COMT-Met are not different from WT at baseline, they may nonetheless have an unutilised higher cognitive reserve that is lacking in WTs. Thus, COMT-Met mice are protected from the negative effects of stress on the two-step task (Figure 64 left).

This is similar to Otto et al.'s argument that individual WM capacity, which is closely related to general cognitive ability, protects the individual from the deleterious effects of the stress response i.e. they were less impaired by acute stress (cold pressure test) on model-based decision making (R. A. Otto et al. 2013d). They found that that cortisol reactivity hampered the model-based choices in low, but not high, WM capacity individuals.

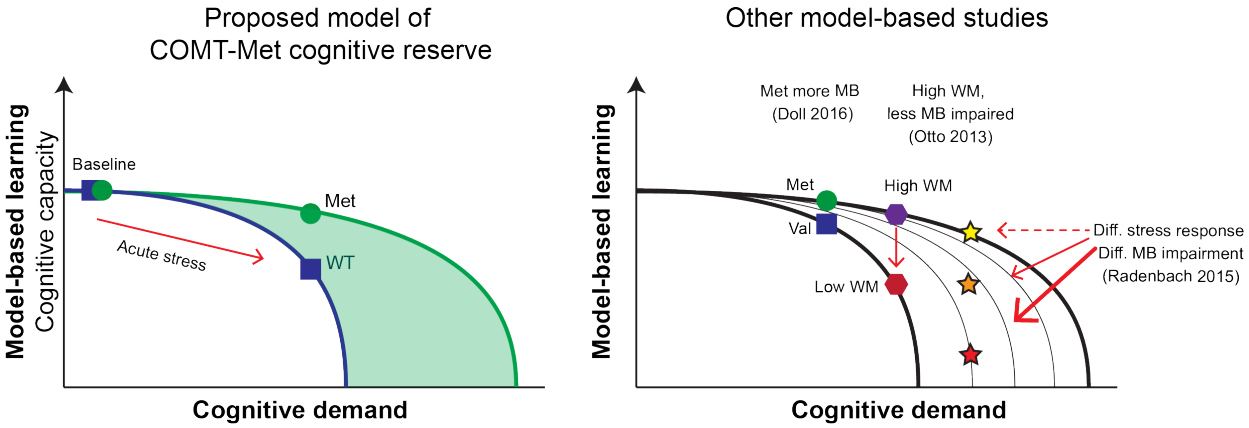


Figure 64: Left: Proposed model of cognitive reserve in which individuals with greater processing capacity are less burdened by the computational expense of model-based estimation and thus, less impacted by the negative effects of stress. Right: The model is similar to Otto et al.'s study of high WM capacity subjects were less impaired by stress on model-based RL than low WM capacity individuals (A. R. Otto et al. 2013b). A model of cognitive reserve may also explain Doll et al.'s finding that Met homozygotes were more model-based than Val carriers (Doll et al. 2016), because the human two-step task in demanding enough to draw on the cognitive reserve of the Met homozygotes. Radenbach et al.'s finds that model-based impairment is correlated with individual stress response (Radenbach et al. 2015).

The idea of *cognitive reserve* also congregates with the *resource modulation hypothesis* and studies of cognition during old age and COMT (Figure 65 right). The resource modulation hypothesis argues that the effects of common genetic variations on brain and cognition e.g. COMT may become stronger in old age due to decline in brain integrity, which amplify to between-person differences (under the assumption that the relationship between brain resources and cognitive performance is non-linear). Indeed, several studies observe the effects of COMT genotype on cognitive performance are magnified in childhood and old age (Lindenberger et al. 2008; Nagel et al. 2008), Figure 65 left.

To test the model of cognitive reserve the cognitive demand i.e. difficulty of the task could parametrically modulated e.g. by changing the contrast between reward and transitions probabilities (0.9/0.1←0.8/0.2→0.7/.03), which may reveal the cognitive reserve at baseline corroborate the human findings (Doll et al. 2016), Figure 64 right). Alternatively, one could increase the intensity of stress e.g. by using multiple stressors or modulate the physiological stress response by amplifying with injecting corticosterone instead of just saline or inhibit by

glucocorticoid receptor antagonists

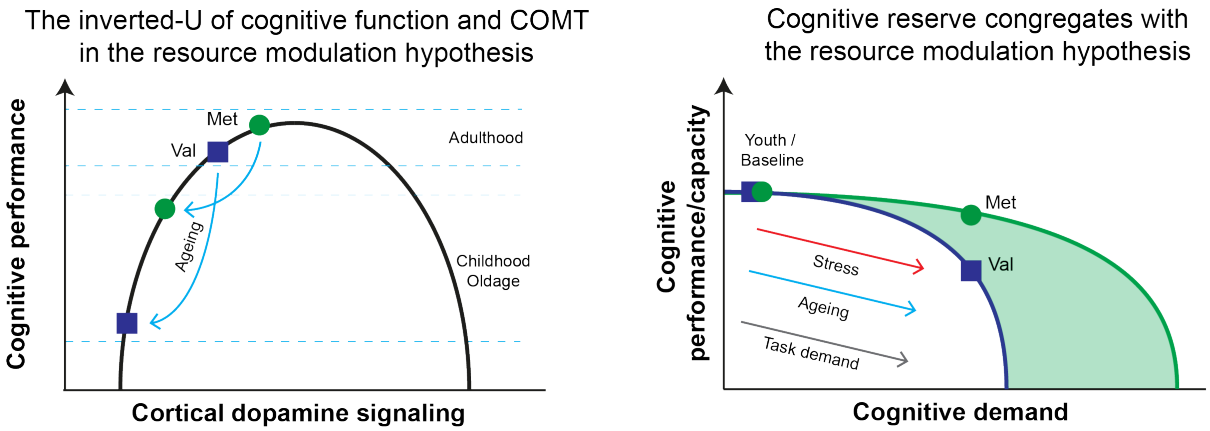


Figure 65: The resource modulation hypothesis and the cognitive reserve hypothesis. Left: Val homozygotes subjects have a greater decline in cognitive function in old age, than Met homozygotes in line with the resource modulation hypothesis that loss of brain integrity due to aging increases the genetic variation of brain function. Image adapted from (Nagel et al. 2008). Right: The resource modulation hypothesis fits with the cognitive reserve hypothesis, that Met homozygotes have a cognitive reserve, that makes the effects of stress, ageing, or task demands less detrimental than their Val counterparts.

Another human two-step study found that the individual stress response correlated with the degree of impairment in model-based control (Radenbach et al. 2015). Furthermore, that stronger cortisol-reactivity was related to a lower degree of model-based control, while higher subjective reports of arousal were associated with a higher frequency of model-based choices. This does not conflict with the observed results or the model of cognitive reserve, if the cognitive resources interact with the stress and arousal i.e. an individual with higher cognitive capacity find the task less stressful (Figure 64 right).

This points back towards the Inverted-U of arousal and performance (Yerkes and Dodson 1908), but more interestingly towards the two additional directions: HPA-axis activation and persistent trait factor stress reactivity. Glucocorticoid receptors in the PFC appear to regulate the stress-evoked DA efflux and associated impairments of cognitive functions (Butts and Phillips 2013), which would affect model-based learning. DA in the nucleus accumbens core (part of the ventral striatum) is also elevated in response to stress during appetitive tasks (Inglis and Moghaddam

1999; Tidey and Miczek 1996), and this has been found to be decreased by corticotropin-releasing factor in the ventral tegmental area that projects to the ventral striatum (Wanat et al. 2013). Under the assumption that model-free learning depends on striatal dopamine, this in principle could then enhance model-free choices and therefore may explain the impairment in model-based learning and enhancement of model-free in the WT mice.

A COMT-genotype modulated HPA-axis activation could contribute to the observed differences observed in the mice. Studies of acute stress in children, but not adults, have found COMT-dependent cortisol responses (Armbruster et al. 2011; Buckert et al. 2012; Mueller et al. 2012; Qin et al. 2009). Female COMT-KO mice had an increased corticosterone response to acute (2 hrs restraint) stress (but not chronic) and increased anxiety (light–dark test) compared with wildtype (C57BL6/J) mice (Desbonnet et al. 2012). An unpublished pilot study from our laboratory found a numerical increase in blood corticosterone levels in male COMT-Met mice, compared to WT, at baseline: An i.p. injection of saline given 30 min prior to corticosterone measurement increased corticosterone levels in both genotypes but abolished the genotype difference observed in the non-stressed condition (A. Huber, personal communication, August, 2016). This suggests that corticosterone could interact with the COMT-dependent cortical functions during stress. This could be incorporated into a future study by measuring corticosterone from the tail vein of the mice in relation to the sessions.

Future studies of COMT-mice will enable us to better understand the underlying neurobiology of different learning strategies while further studies of acute stress the adaptive mechanisms employed in maximising reward while adapting to an ever-changing environment.

6. General discussion

The aim of this thesis was to investigate the effect of COMT genotype, and acute stress, on model-based and model-free RL. The rationale for this research is based on the following: That COMT genotype modulates cortical dopamine and function, which model-based decision making relies on, and that cortical function and model-based learning are both impaired by acute stress.

6.1. Water and Sucrose Reinforcement (Chapter 3 and 4)

Before studying the effect of injection and COMT genotype, it was necessary to adapt the reinforcement protocol from water to sucrose. Several findings emerged from this.

First, sucrose-reinforced subjects performed fewer trials than water reinforced subjects. This was prominent in the first hour, where the water-reinforced mice performed a very high number of trials. In contrast, the sucrose-reinforced mice did have an initially high trial activity, but after only half an hour established a low, but stable, trial rate. We extended the length of sessions and modified the task slightly to compensate for the fewer trials per session. Across sessions, we observed an increase in model-based influence.

Secondly, coinciding with the initial drop of activity in the sessions of sucrose-reinforced mice, we observed a decrease in model-based RL, consistent with satiation (B. W. Balleine 1992b; B. W. Balleine and Dickinson 1998; Colwill and Rescorla 1985). Moreover, during the whole session, we also observed a decline in model-free decision-making, which I interpret as being due to modulation of the mesolimbic dopamine (and magnitude of RPE) by the physiological state of the animal in order to prioritise actions that maintain homoeostasis (Papageorgiou et al. 2016). Surprisingly, this was not observed in the water-reinforced mice, which could be attributed to the fact that they had a dry food pellet in the testing box to

maintain motivation for work for water. These findings suggest that motivation and reward value do not show the same pattern of satiation for water as for sucrose reward.

Third, there were only subtle differences in the performance of the mice between the two protocols: the sucrose-reinforced subjects had numerically, but non-significantly, lower model- based RL than the water-reinforced animals.

However, the results regarding the difference between water and sucrose reinforcement must be interpreted with caution, in addition to the reinforcer, the two cohorts were run separately, in different locations, and thus varied in many other factors.

Nonetheless, we are not the first to find differences in behaviour dependent on the identity of the reinforcer used. Tabbara et al. observed that water deprivation, but not sucrose concentration increased the acquisition of Pavlovian conditioning, arguing that water deprivation is a special, highly-motivating state (Tabbara et al. 2016). Furthermore, studies of the RPEs encoded by mesolimbic dopamine were modulated by the physiological state of the animal (pre-fed/not pre- fed) in an appetitive task (Papageorgiou et al. 2016). In addition, maltose was observed to be more reinforcing than sucrose in rats despite equivalent caloric value, suggesting that caloric content and physiological state alone do not play a role, but that other factor are also important, such as flavour (Roop et al. 2002).

These findings highlight the need to control for within and across session changes in RL, but also how appetitive reinforcers used in behavioural experiments might subtly, but dynamically, influence decision making, possibly due to a modulation of the physiological state of the animal, which should be taken into consideration when selecting a reinfocer for a given study and comparing data from tasks using different reinforcers.

6.2. Acute stress influences RL (chapter 4 and 5)

As part of a larger pharmacological study, it was possible to examine the subtle effects of a commonly overlooked stressor in pharmacological studies of behaviour in rodents: injection stress (Drude et al. 2011). When wild-type mice were given an i.p. injection of a commonly-used vehicle (cyclodextrin) 15 min before testing, we observed distinct effects on behaviour.

Similar to humans exposed to acute physical or psychosocial stress, I observed in wild-type mice a decreased model-based decision making and shifted toward a more model-free decision- making (A. R. Otto et al. 2013b; C. Radenbach et al. 2015a). The same pattern was observed in another group of wild-type (COMT-WT) mice injection with saline 10 min before testing.

This adds to the evidence that both humans and rodents when acutely stressed are impaired in goal-direction decision making and fall back on habits (Braun and Hauber 2013; Dias-Ferreira et al. 2009; Christoph Radenbach et al. 2015b; Lars Schwabe and Wolf 2011). It would also be of interest to complement these findings in a study employing an acute psychosocial or naturalistic stressor, such the odour of a predator, to see if similar results are obtained.

When the mice were injected 90 min before testing with vehicle, they showed a similar pattern of model-free increase and model-based decrease at levels numerically intermediate to 15 min- prior injection group and the non-injection controls. This suggests a time-since-injection effect, or a dose-dependent effect of stress, assuming the level of stress had declined since injection. However, we do not have glucocorticosteroid measurements to assert this.

Therefore, a further exploration of the effect of acute stress could include parametrically

modulating the intensity of the stress. This could be achieved by adding an additional stressor or adding exogenous glucocorticosteroid to the injection. Alternatively, one could measure glucocorticosteroid levels during the session of each animal to explore the relationship between glucocorticosteroid levels and decision making.

The impact of injection stress needs to be taken into account in pharmacological studies where it may mask a drug's effect on behaviour – in the absence of a non-injected control, an apparent drug effect may in fact be confounded by the effect of acute stress.

6.3. COMT-genotype influences RL (chapter 5)

The COMT-polymorphism influences PFC dopamine and cognitive performance. A human two-step study found that the Met allele dose-dependently increased model-based (but not model- free) decision-making (Doll et al. 2016).

I, therefore, hypothesised that COMT-Met mice would be more model-based based than their COMT-WT littermates at baseline. However, I did not find such a difference in performance or decision-making style. There are a number of reasons for the inconsistency between my findings and those reported earlier. Firstly, there are several differences between the human and rodent versions of the two-step task (the rodent version had one rather than two choices, and had reward probabilities change in blocks rather than after each trial); Thus, the rodent task may not be completely analogous to the human version. Similar to human cognitive studies of the COMT polymorphism, it is possible that the lack of genotype differences may be due to relative differences in task requirement and difficulty between the human and mouse versions. Buckert et al. (2012) observes genotype differences on a working memory task on the intermediate level (2- back), but not at the easy (1-back) or difficult (3-back) level (Buckert et al. 2012). Secondly, mice are not humans

and there are important species differences in a number of aspect of cognitive function, particularly that dependent on the PFC. Thus, the behaviours observed in mice may not directly reflect the cognitive processes underlying human model-based RL. Thirdly, it is possible a genotype effect may be overlooked due to the limited sample size (although the use of littermate controls should limit variability between COMT-Met and WT groups in a way that is not possible in human studies).

These data therefore suggest that there is no effect of COMT genotype on model-based and model-free RL at baseline, although these results should be treated with some caution, given the caveats outlined above. In the future, it would be of interest to investigate the impact of parametrically modulating cognitive demand by changing the difficulty of the task, to see whether genotype differences emerge under particular task conditions.

6.4. COMT-genotype may modulate the effect of acute stress on RL (chapter 5)

The PFC is sensitive to acute stress, which can lead to prefrontal dopaminergic dysfunction and impairment of cognitive function (A. F. Arnsten 2009a; A. F. T. Arnsten 2015). Acute stress negatively impacts working memory in human Met subjects, whilst improving it in Val subjects (Buckert et al. 2012; Qin et al. 2009). This interaction between genotype and stress has also been observed for PFC activation during working memory: stress increased working memory-related activation in the PFC in Val carriers but decreased it in Met homozygotes (Qin et al. 2009).

Furthermore, a PET study found less PFC dopamine release in Met carriers compared to Val homozygotes in response to acute stress (Hernaus et al. 2013).

These data suggest that the COMT polymorphism modulates the effect of stress on

cortical dopamine and working memory, both of which are involved in model-based learning, and that acute stress might be beneficial to working memory performance in Val carriers, but detrimental in Met carriers. However, in contrast, another study observed that having a high working memory is protective against the stress-induced impairment of model-based RL (R. A. Otto et al.
2013d). Since the Met allele is predictive of higher working memory capacity (at least in some studies, as outlined in the Introduction), this would predict that the Met allele should protect against some of the negative effects of acute stress.

Although conflicting, these relationships are not mutually exclusive. High working memory at baseline may be associated with other working memory or PFC dopamine-independent neuronal resources involved in model-based RL that may lessen the impact of acute stress. Consistent with this latter proposal, despite no discernible difference on performance and decision-making style at baseline in the COMT mice, I hypothesised that COMT-Met mice might have a higher cognitive reserve than WTs, that though unutilised at baseline, would appear when the cognitive demand increase under stress.

The COMT-WT mice showed the pattern previously observed in wild-type (C57BL/6) mice: an increase in trial rate, and a decrease in model-based and increase in model-free RL following acute stress (injection).

In contrast, the Met mice did not increase their trial rate after injection, but had an increased tendency to repeat choices after common transitions. Indeed, there was an interaction between genotype and stress on the tendency to repeat choices following transitions, and independent of outcome and transition (the latter must be interpreted with caution, as it was only $P<0.1$). The WT mice increased their trial activity in a manner

whereby they increased repeating choices independent of the outcome or transition in the previous trial. Thus, WT were responding in a more habitual way disregarding the structure of the task, whereas the Met mice became more aware of the task structure after an injection.

Most strikingly, COMT-Met mice were much less affected by injection stress and did not show significant changes in either model-based or model-free learning. However, when single-subject fitted values from the RL model were calculated and analysed by RM-ANOVA, there were neither main nor interactive effects of genotype and injection stress, meaning that the numerical differences between the groups must be treated with caution. However, it is also important to note that there is a considerable loss of power when fitting the RL model on 6 sessions per subject per condition rather than the 48 sessions of the group per condition, and taking full advantage of the multilevel modelling approach and the statistical approach of permutation the whole dataset (256 sessions). Therefore, the RM-ANOVA is a rather conservative analysis and should not definitively be taken as evidence for lack of differences between the groups.

An improvement of this study would be to increase sample size, and develop a permutationtest for the RL model to study interactions of within- and between-subject factors. This would require a high-performance computer cluster (similar to that used for the analysis of fMRI data).

Taken together, the data suggest that the Met mice were less effected by stress. I hypothesise that this is because they have a higher cognitive reserve to cope with the negative effects of injection stress on performance. To study this, I propose modulating the cognitive demand of the task by increasing and decreasing the difficulty e.g. by changing

the contrast between probabilities. This would enable me to identify not only the presence but also the magnitude of the cognitive reserve. In addition, it would be of interest to modulate the identity and manner of the stressor to future study the genotype-stress interaction.

Taken together, these findings suggest that subtle effects of genetic variation might be amplified under conditions of stress, providing a potential explanation for some of the controversies seen in the human literature linking COMT genotype and cognitive function.

7. Appendix

7.1. Appendix to Chapter 1

Chapter one has no additional information.

7.2. Appendix to Chapter 2

7.2.1. Development of the COMT mouse model

The COMT-Met mice were developed by Jingshan Chen and Daniel Weinberger at the National Institute of Mental Health and the Lieber Institute for Brain Development in the United States. In the COMT-Met mice the native COMT amino acid (Leu[148]), which is equivalent to the human Val[158]Met locus, was replaced with methionine by a PCR-based strategy (see Figure 66). COMT-Met littermates, COMT-WT, retain the Leu amino acid.

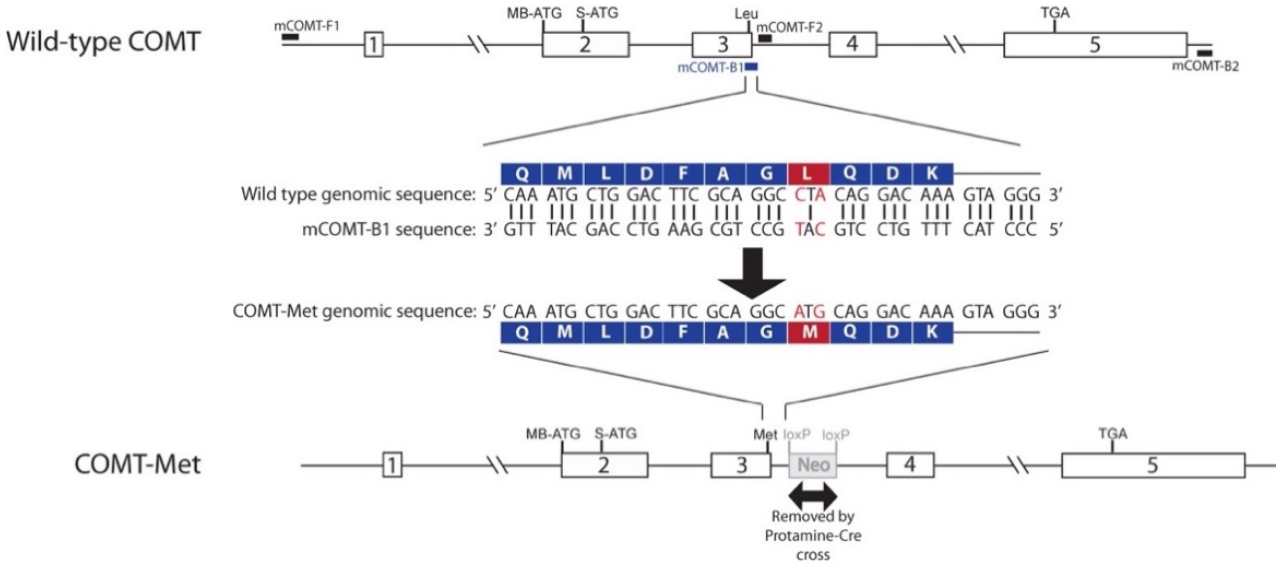


Figure 66: Development of COMT-Met mice. Site-directed mutagenesis was used to generate mouse COMT genomic DNA containing the Met allele. The mCOMT-B1 primer introduced the Met allele into the mouse COMT gene replacing the native Leu (mismatched bases in red). The DNA fragment was inserted into a construct containing a neomycin selection cassette flanked by *loxP* sites (PGK-neo) in the intron between exon 3 and 4. That was used to generate mice on a mixed 129sv-C57BL/6 background according to standard techniques (Crawley 2007) and backcrossed for three generations with C57BL/6 mice. To remove the neomycin selection cassette from the mice, they were crossed with a line of mice expressing Cre-recombinase under the control of the Protamine promoter (Jackson Laboratory) to remove the neomycin selection cassette. Mice were then further backcrossed onto a C57BL/6 background for an additional 8 generations. Image and legend from (Barkus et al. 2016).

All procedures in the US were approved by the National Institute of Mental Health Animal Care and Use Committee, U.S. and followed the National Institutes of Health Using Animals in Intramural Research, U.S. guidelines. They were transferred to the UK, where all procedures were carried out in agreement with the Animals (Scientific Procedures) Act 1986 and Home Office guidelines.

7.2.2. COMT genotyping

COMT genotype was confirmed by PCR on genomic DNA extracted from ear clips. This was done using a restriction fragment length polymorphism assay: a fragment of the COMT gene was amplified using the primers Met-Mouse-Geno-F (`5′ GCC AAA TCA TGG ATG CAG TGA TTC 3′`) and Met-Mouse-Geno-R (`5′ CTT TGG GGC TTT GGA GAG GAA 3′`). The resulting 471 bp amplicon was digested with the restriction enzyme *StuI*. The enzyme recognises `AGG^CCT` and cleaves the amplicon into 244 bp and 227 bp fragments in Leu mice. The presence of the Met allele removes the *StuI* site. When visualised on agarose gels Met homozygotes show only a 471 bp band, Leu homozygotes show only 244 bp and 227 bp bands, and heterozygotes show all three bands. (Huber 2015)

7.2.3. Counterbalancing of injections for SUCROSE 2

Day	1	2	3	4	5	6	7	8	8	10
Cycle	1st						2nd…			
Side		Right		Left		Right		Left		
Mouse 1	None	Veh-15	None	Tolc-90	None	GBR-15	None	Tolc-90	None	Veh-90
Mouse 2	None	Tolc-90	None	Veh-90	None	GBR-15	None	GBR-15	None	Veh-15
Mouse 3	None	GBR-15	None	Veh-15	None	Tolc-90	None	Veh-90	None	GBR-15
Mouse 4	None	Veh-90	None	GBR-15	None	Tolc-90	None	Tolc-90	None	GBR-15
Mouse 5	None	Tolc-90	None	GBR-15	None	Veh-15	None	GBR-15	None	Tolc-90
Mouse 6	None	GBR-15	None	Tolc-90	None	Veh-90	None	Veh-15	None	Tolc-90
Mouse 7	None	Veh-15	None	Tolc-90	None	GBR-15	None	Tolc-90	None	Veh-90
Mouse 8	None	Tolc-90	None	GBR-15	None	Veh-90	None	GBR-15	None	Tolc-90

Table 19: Overview of counterbalancing. The ratio between injections during each cycle was 2 Tolcapone, 2 GBR, 1 Veh-15, 1 Veh-90, 6 No-injection

7.2.4. Picture of testing box wall



Figure 67: Back board with nose poke holes with a two pound coin for size.

7.3. Appendix to Chapter 3

Chapter 3 contains no additional information.

7.4. Appendix to Chapter 4

7.4.1. Diet and sucrose contribution during testing

Diet size and content has been linked to cognitive performance in both humans and rodents (T. Murphy et al. 2014) and can be modulated by COMT-genotype (Witte et al. 2010). As the mice increased their activity and earned more rewards, the dietary energy contribution from sucrose increased (Figure 68). To counter this and increase motivation the reward size was stepwise lowered to 4 µl. During testing the mice were fed a 16% protein standard chow diet, kept stable at 85-90% of their full fed weight and had a Body Score of 3. The sucrose contribution to dietary energy was limited to approximately 10% on testing days and lower during the whole testing period (Table 20). Thus, they did not become obese or high-sucrose diet. Therefore, despite the link between diet and cognition, this is not considered further here.

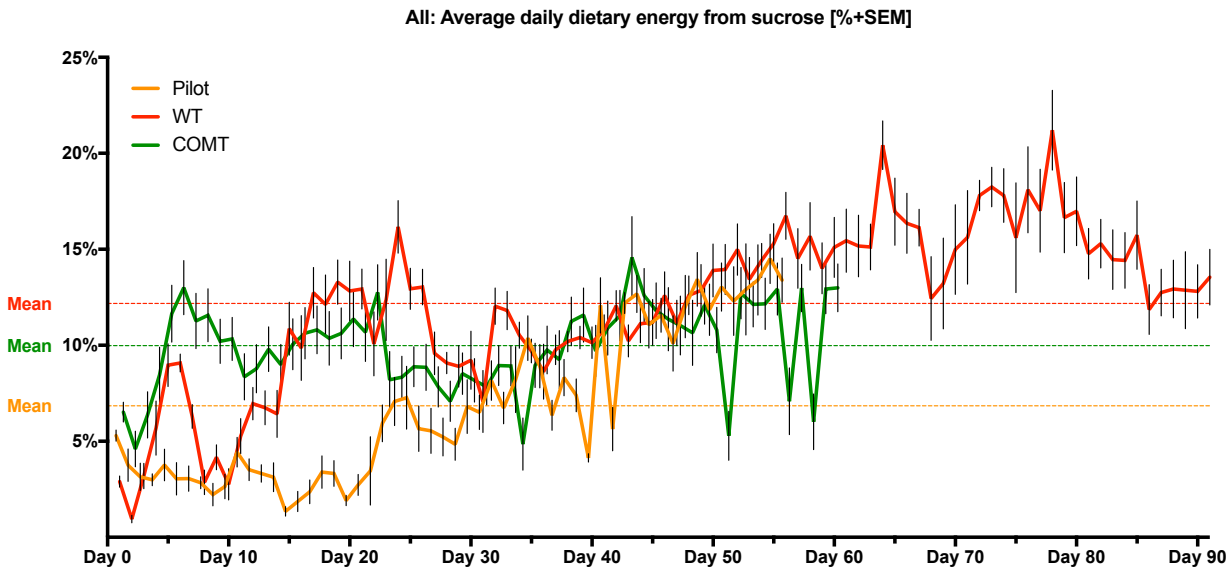
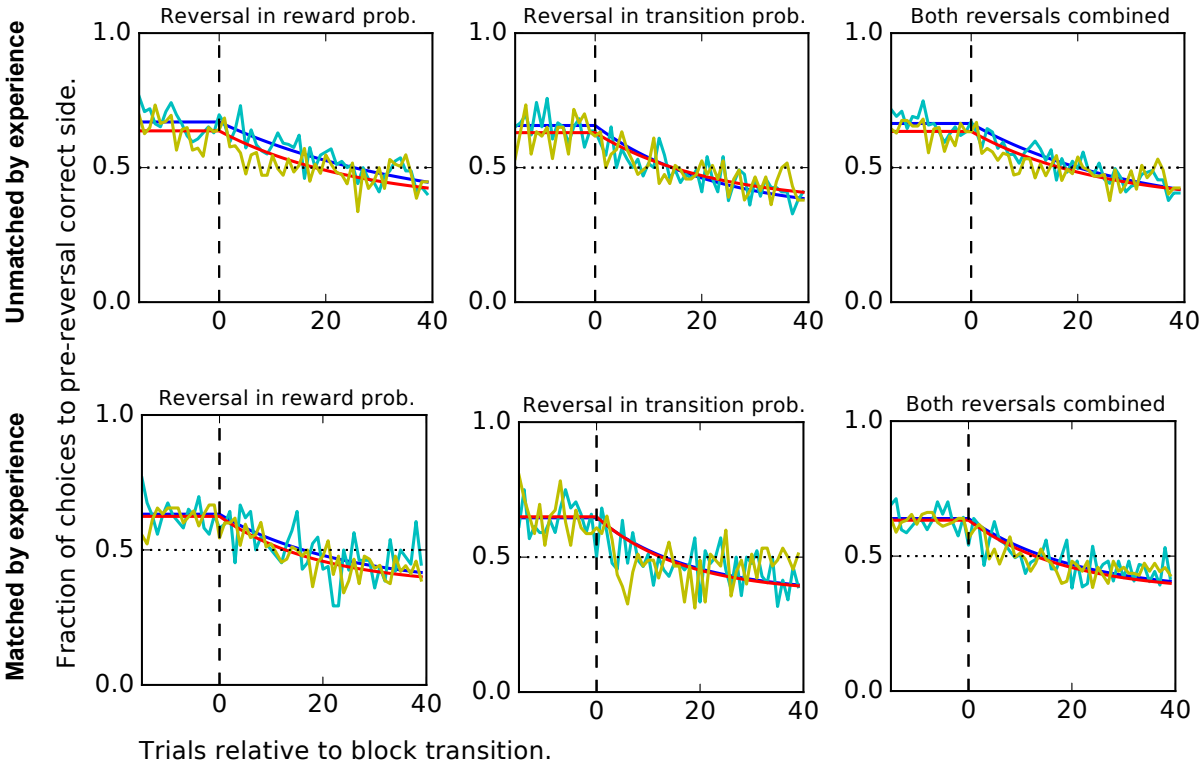


Figure 68: Daily dietary energy from sucrose in the cohorts during testing. As activity rose during testing and sucrose consumption increased the reward size was stepwise reduced.

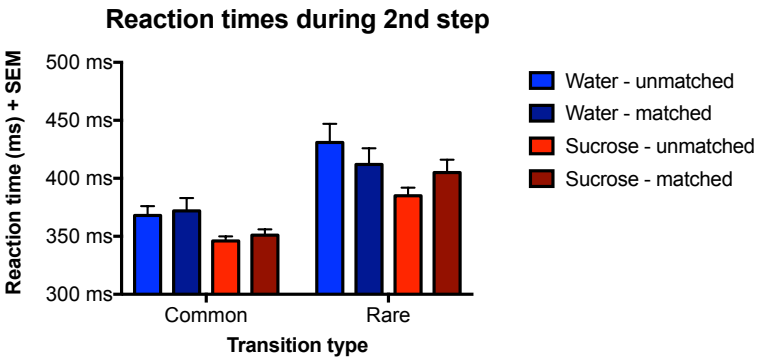
Cohort	Testing	Testing period	Weight	% free-feeding Weight during testing	Mass of chow given (g)	Energy from chow, test days	Energy from sucrose, test days	Energy from sucrose, test days	Energy from sucrose, all days
	Days	Days	g	%	g	kJ	kJ	%	%
WATER 1	39	42	24.8±1.3	94±8	n/k	n/a	n/a	n/a	n/a
WATER 2	40	45	23.0±1.3	95	n/k	n/a	n/a	n/a	n/a
SUCROSE 1	57	118*	25.5±1.1	88±2	2.6±0.3	32.2±3.3	2.34±1.39	6.7±4.1	3.2±2.0
SUCROSE 2	91	102	20.2±.13	88±4	2.0±0.3	25.2±3.8	3.44±1.46	12.2±4.1	10.9±3.7
COMT	60	68	30.6±1.9	85±3	2.4±0.4	30.2±5.0	3.40±1.92	10±6	3.0±1.7

Table 20: Dietary energy during the studies. All values mean ± SD. *After the 10th day of training the Pilot cohort had a 30 day free feeding period without testing. n/a = not applicable; n/k = not known

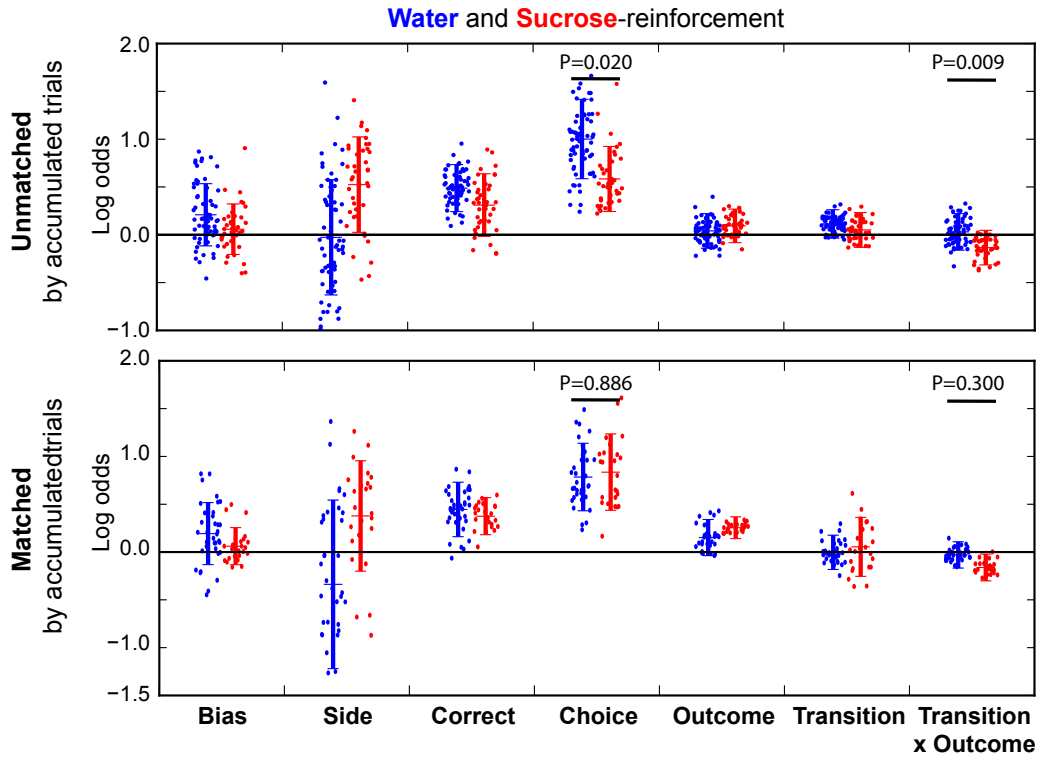
7.4.2. Permutation testing of Water versus Sucrose



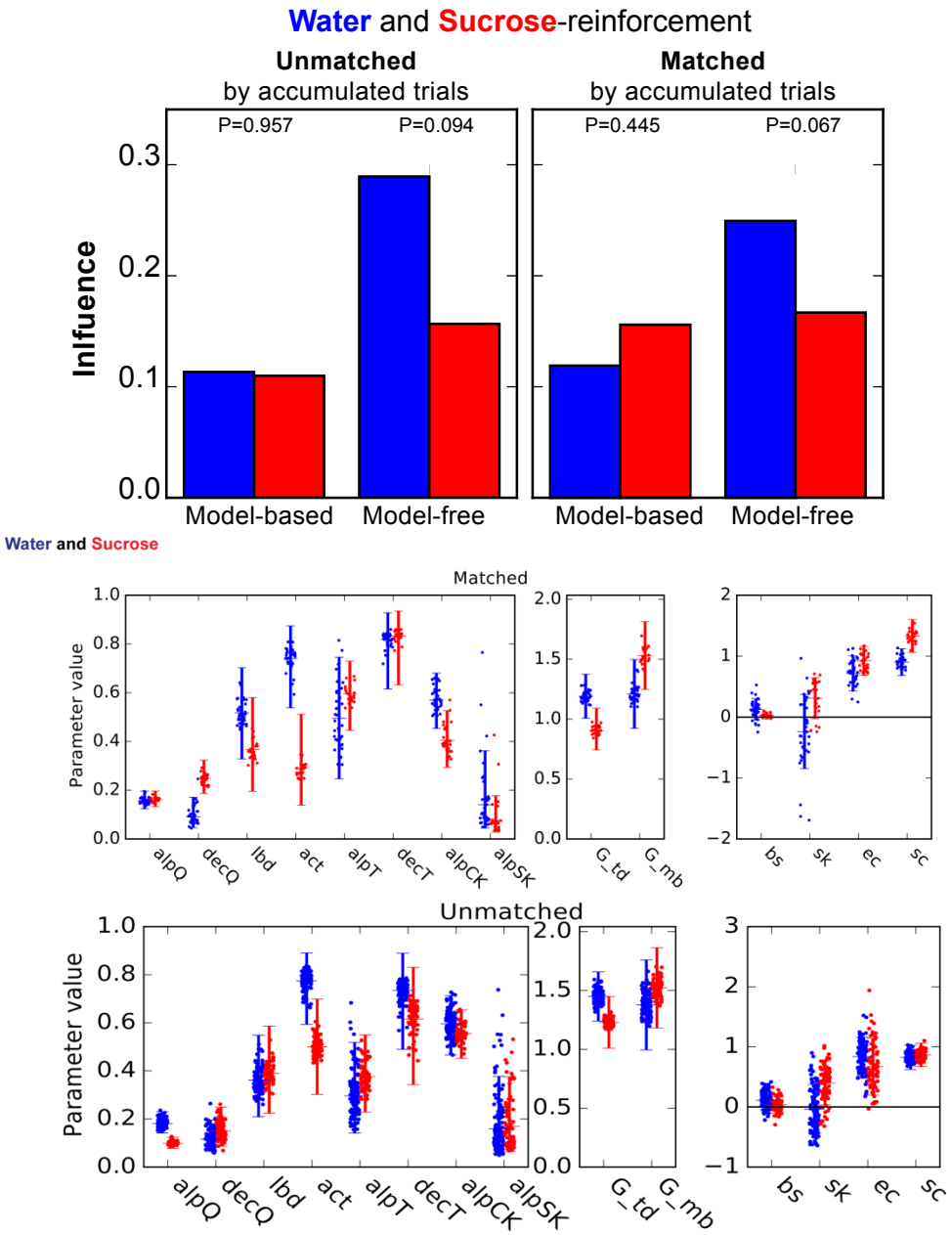
Unmatched by experience (accum. trials)	Matched by experience
Water 1+2 (Days 30-35)	Water 1 (Days 28-30) + Water 2 (days 25-27)
Average block end choice probability: 0.671	Average block end choice probability: 0.635
Reward probability reversal, tau: 36.563	Reward probability reversal, tau: 23.591
Trans. probability reversal, tau: 22.086	Trans. probability reversal, tau: 17.831
Combined reversals, tau: 29.770	Combined reversals, tau: 20.480
Sucrose 2 (Days 30,31,32,33,35)	Sucrose 2 (Days 37-39)
Average block end choice probability: 0.634	Average block end choice probability: 0.634
Reward probability reversal, tau: 25.569	Reward probability reversal, tau: 19.367
Trans. probability reversal, tau: 21.734	Trans. probability reversal, tau: 16.673
Combined reversals, tau: 24.196	Combined reversals, tau: 18.637
Comparison by permutations across subjects	Comparison by permutations across subjects
Block end choice probability P value: 0.060	Block end choice probability P value: 0.980
All reversals tau P value : 0.554	All reversals tau P value : 0.789
Reward prob. reversal tau P value: 0.425	Reward prob. reversal tau P value: 0.663
Trans. Prob. reversal tau P value: 0.937	Trans. Prob. reversal tau P value: 0.894



Unmatched by experience (accum. trials)	Matched by experience (accum. trials)
Permutations cross subjects Common RT P value:0.0500 Rare RT P value:0.0260 Diff. RT P value:0.0740	Permutations cross subjects Common RT P value:0.1250 Rare RT P value:0.7160 Diff. RT P value:0.3640



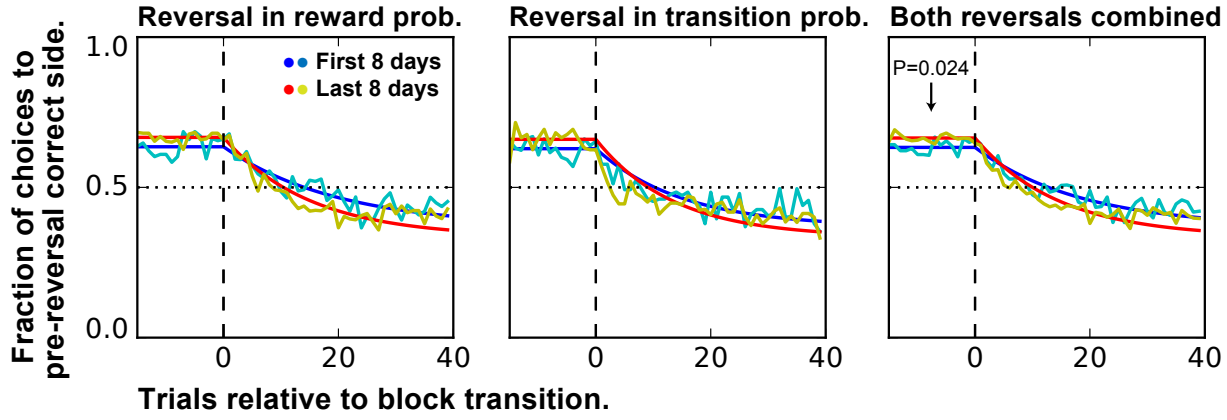
Unmatched by average accumulated trials:	Matched by average accumulated trials:
Water1+2 (Days 30-35) versus Sucrose2 (Days 30,31,32,33,35) P values: bias : 0.187 side : 0.022 correct : 0.112 choice : 0.020 outcome : 0.498 trans_CR : 0.333 trCR_x_out: 0.009	Water1 (Days 28-30) + Water2 (days 25-27) vs Sucrose 2 (Days 37-39) P values (1000 permutations): bias : 0.330 side : 0.090 correct : 0.375 choice : 0.886 outcome : 0.332 trans_CR : 0.772 trCR_x_out: 0.300



Unmatched by average accumulated trials:	Matched by average accumulated trials:
Water1+2 (Days 30-35) versus	Water1 (Days 28-30) + Water2 (days 25-27)
Sucrose2 (Days 30,31,32,33,35)	vs Sucrose 2 (Days 37-39)
P values (1000 permutations):	P values (1000 permutations):
alpQ : 0.211	alpQ : 0.724
decQ : 0.524	decQ : 0.033
lbd : 0.874	lbd : 0.623
act : 0.310	act : 0.021
alpT : 0.645	alpT : 0.522
decT : 0.202	decT : 0.986
alpCK : 0.322	alpCK : 0.203
alpSK : 0.353	alpSK : 0.032
G_td : 1.000	G_td : 0.197
G_mb : 0.907	G_mb : 0.449
bs : 0.232	bs : 0.200
sk : 0.018	sk : 0.049
ec : 0.265	ec : 0.480
sc : 0.382	sc : 0.076
Model-based influence: 0.957	Model-based influence: 0.475
Model-free influence : 0.094	Model-free influence : 0.062

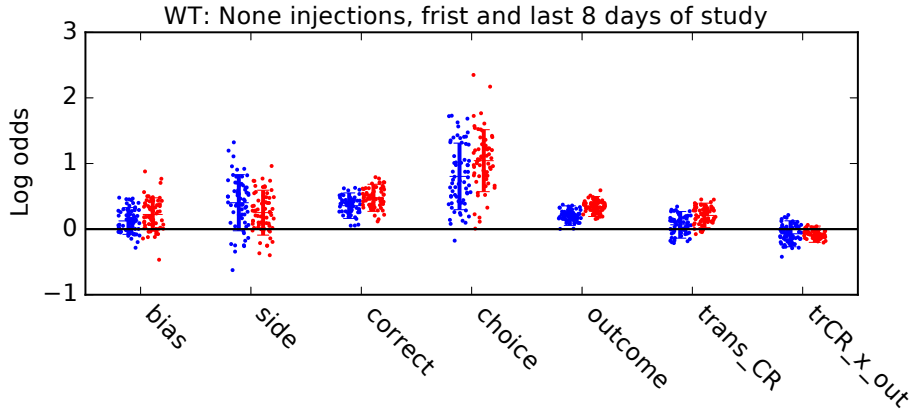
7.5. Appendix to Chapter 5

7.5.1. Across session changes



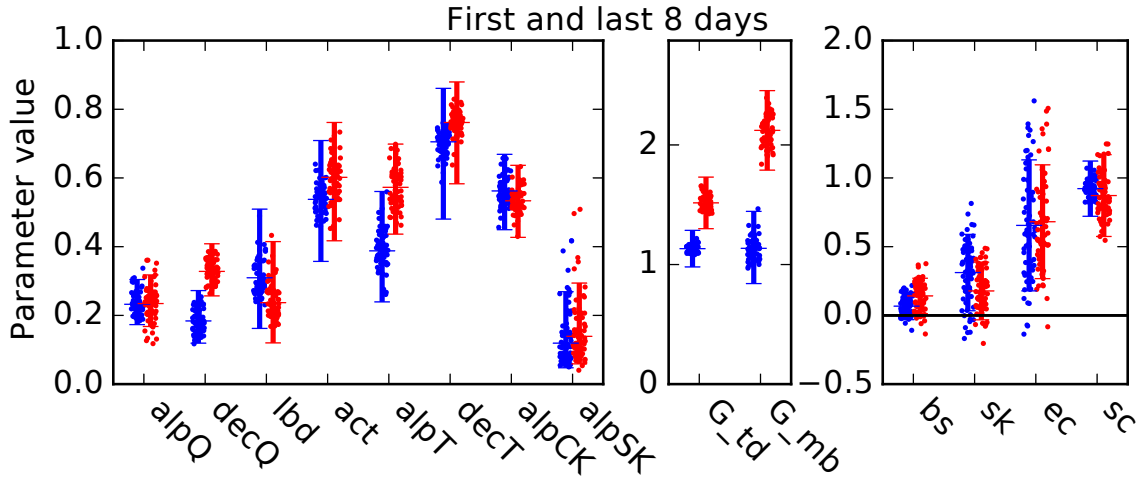
*Within subject permutations	First 8 days	Last 8 days	*P=
Average block end choice probability:	0.632	0.666	0.018
Reward probability reversal, tau:	19.98	15.01	0.099
Trans. probability reversal, tau:	15.08	13.51	0.690
Combined reversals, tau:	18.39	14.55	0.126

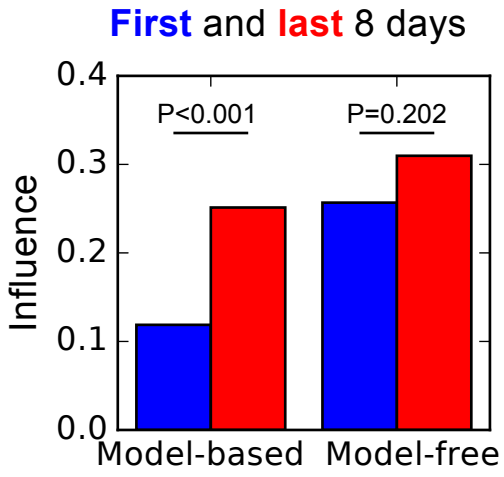
WT: First 8 days, last 8 days - Logistic regression model - within subject, without cycles



P values (1000 permutations):
bias : 0.010
side : 0.039
correct : 0.043
choice : 0.000
outcome : 0.002
trans_CR : 0.022
trCR_x_out: 0.918

WT: First 8 days, last 8 days – RL model





Within-subject permutation-testing,
P values (1000 permutations):

alpQ	0.783
decQ	0.087
lbd	0.909
act	0.131
alpT	0.018
decT	0.471
alpCK	0.236
alpSK	0.731
G_td	0.075
G_mb	0.013
bs	0.023
sk	0.015
ec	0.475
sc	0.806
Model-based influence	0.000
Model-free influence	0.202

7.5.2. Within session changes

7.5.2.1. Within-session changes with water-reinforcement

The sessions from both water-reinforced cohorts (WATER 1 days 29-39 and WATER 2 days 29-40) were divided into the first hour with the high trial rate and the second hour with the lower trial rate (Figure 69 left). Comparison of these showed a significant difference in mean trial rate (within-subject and -cycle permutation testing: $P<0.001$). The subjects also performed more slowly during the 2nd step in the second half of the session \on both common ($P=0.523$) and rare ($P=0.543$) transitions, though this was not significant by permutation testing within-subject and within-cycles of 2 days.

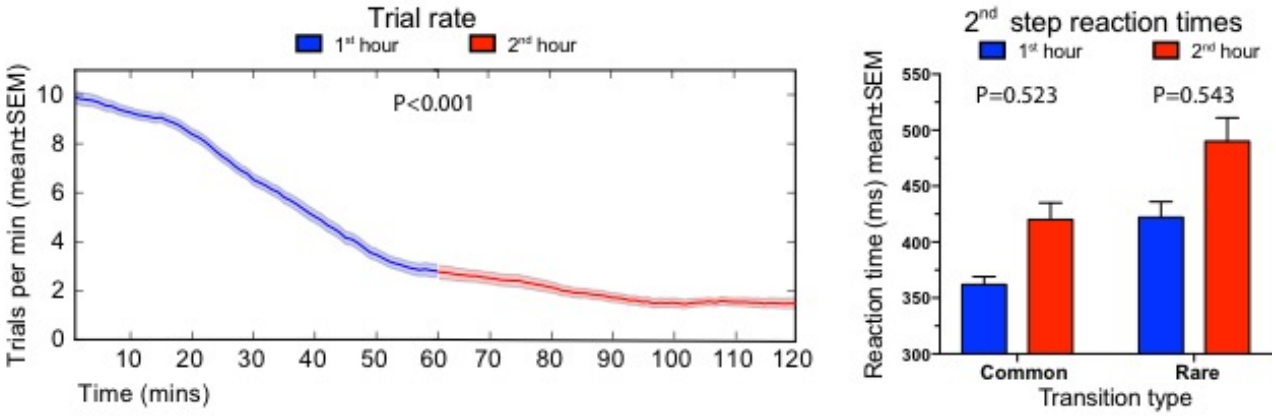


Figure 69: No significant changes in behaviour during the session during water reinforcement apart from trial rate. The first (blue) and second (red) half of the 2 hr sessions from day 29-39 of Water.1 and Water.2 Left) Mean±SEM trial rate (with-subject and cycle permutation-testing: $P<0.001$) Right) 2nd step reaction time presented as mean±SEM of subjects on common ($P=0.523$) and rare ($P=0.543$).

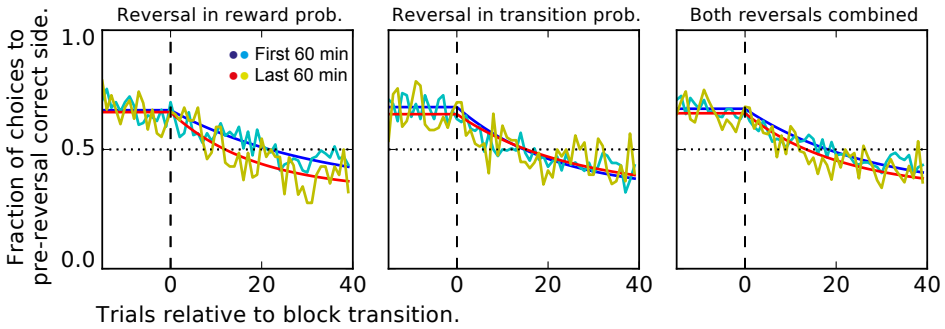
There were no significant changes in block end choice probability (within-subject and cycle permutation-testing: $P=0.696$) or τ after reversals of reward probability ($P=0.138$), transition probability ($P=0.617$) or both combined ($P=0.261$) within sessions. Thus, there were no changes in performance during the sessions according to reversal analysis (Figure 70).

Comparison by logistic regression showed no changes apart from a decreased loading on *choice* (within-subject permutation-testing: $P=0.014$), indicating a decreased tendency to repeat choices independent of previous events e.g. reward or transition (Figure 71). This did not coincide with any

increased tendency to repeat choices following any previous events e.g. reward or transition. Thought, the variation in loading on the predictors of outcome, transition and their interaction was reduced. This suggested no change in model-free or model-based reinforcement learning.

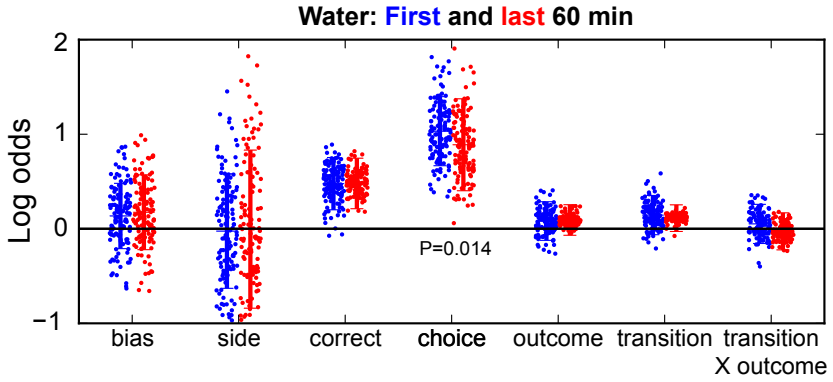
Within the RL model (Figure 72), the side choice bias parameter (*sc*) decreased (within-subject and cycle permutation-testing, *P*=0.019), and the motor-level side perseverance (*alphaSK*) increased (*P*=0.002) together with the model-free estimation weight (*G_td*, *P*=0.012). However, this did not change the model-based (within-subject permutation-testing: *P*=0.992) and model-free (*P*=0.755) influence on decision making.

In conclusion, although the water-reinforced cohorts showed a decline in trial rate during sessions, they had stable performance and model-based and model-free influence.



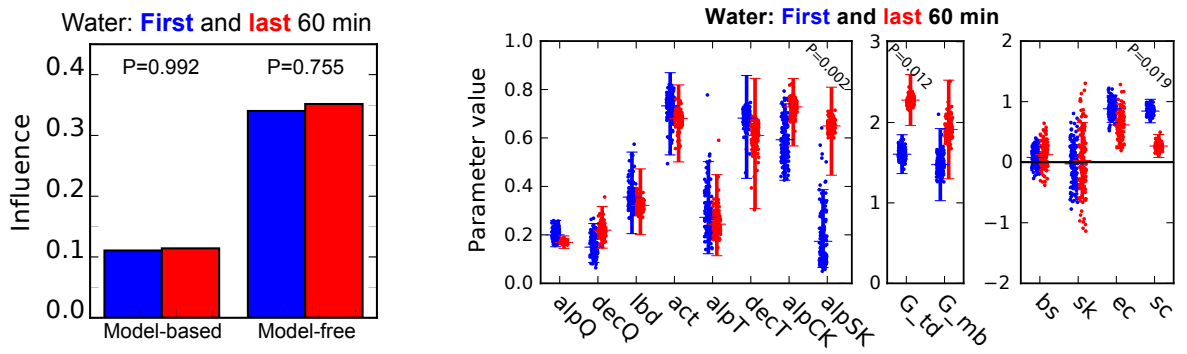
*Within-subject and cycle perm.-testing	First 60 min	Last 60 min	*P=
Average block end choice probability:	0.681	0.674	0.696
Reward probability reversal, tau:	33.563	18.650	0.138
Trans. probability reversal, tau:	21.532	24.780	0.617
Combined reversals, tau:	27.318	21.434	0.261

Figure 70: Reversal analysis of the first and last 60 min from day 29-40 of the water-reinforced cohorts.



P values (1000 permutations): bias : 0.337, side: 0.751, correct: 0.970, choice: 0.014, outcome: 0.822, transition: 0.616, transition-outcome interaction: 0.172

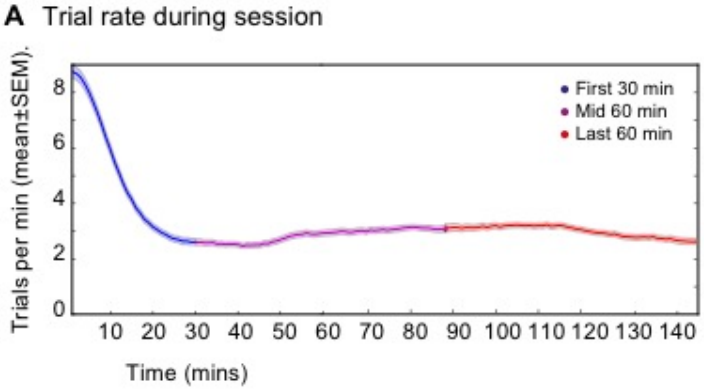
Figure 71: Logistic regression of choice behaviour with water-reinforcement in the first and last 60 min.

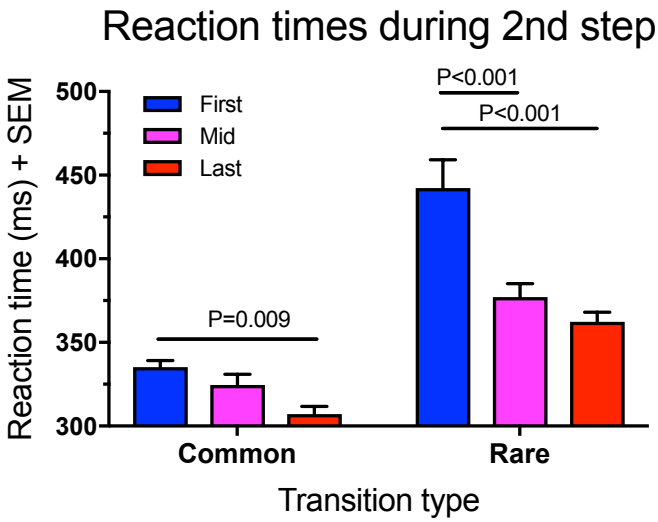


P values (1000 within-subject and cycle permutations): alpQ: 0.495, decQ: 0.169, lbd: 0.603, act: 0.585, alpT: 0.959,decT: 0.666, alpCK: 0.209, alpSK: 0.002, G_td: 0.012, G_mb: 0.364, bs: 0.110, sk: 0.281, ec: 0.118, sc: 0.019; Model-based influence: 0.992 and Model-free influence: 0.755

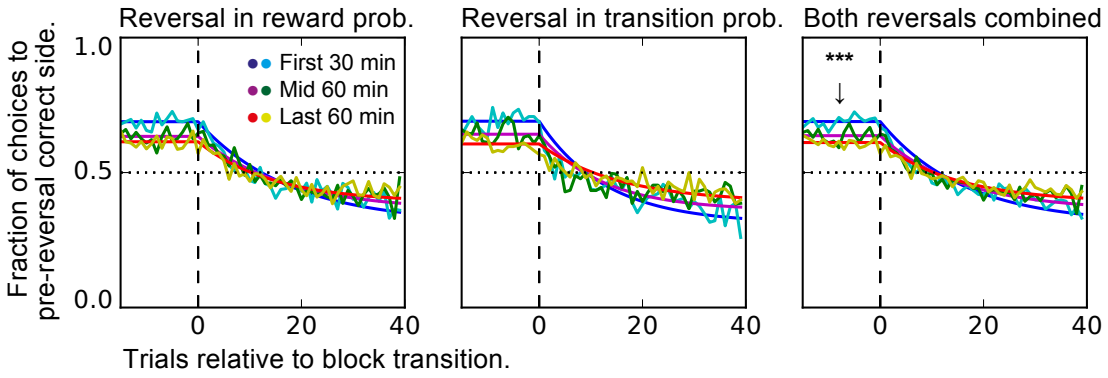
Figure 72: RL model of choice behaviour with water-reinforcement during the first and last 60 min.

3.4.3. Within-session changes with sucrose-reinforcement



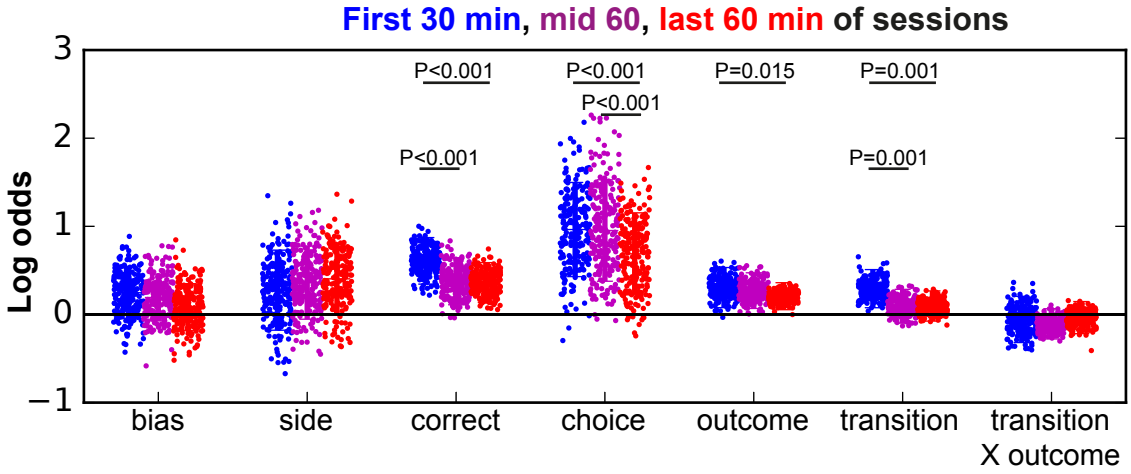


First 30 min vs last 60 min	First 30 min vs Mid 60 min	Mid 60 min vs last 60 min
Trial number difference P value:<0.001	Trial number difference P value:<0.001	Trial number difference P value: 0.001
Common RT P value: 0.484 Rare RT P value: 0.473 Diff. RT P value: 0.468	Common RT P value: 0.494 Rare RT P value: 0.536 Diff. RT P value: 0.543	Common RT P value: 0.483 Rare RT P value: 0.962 Diff. RT P value: 0.510



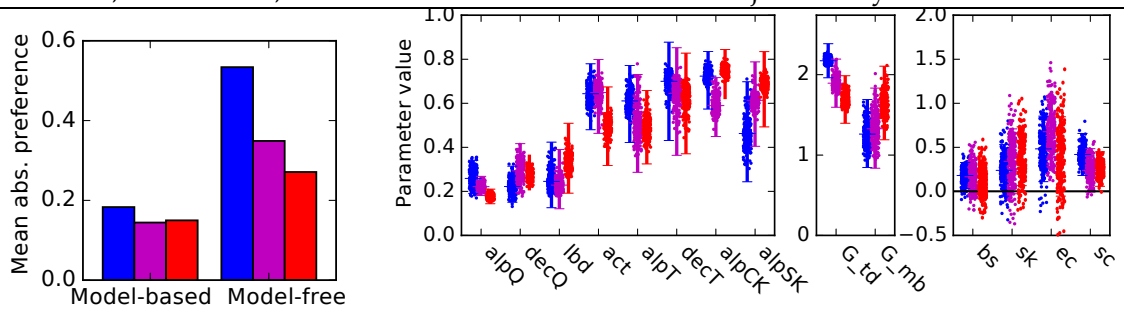
First 30	Mid 60 min	Last 60 min
Average block end choice probability: 0.688 Reward probability reversal, tau: 17.713 Trans. probability reversal, tau: 12.996 Combined reversals, tau: 16.190	Average block end choice probability: 0.640 Reward probability reversal, tau: 16.771 Trans. probability reversal, tau: 12.438 Combined reversals, tau: 15.427	Average block end choice probability: 0.610 Reward probability reversal, tau: 14.473 Trans. probability reversal, tau: 15.524 Combined reversals, tau: 14.828
First 30 min vs last 60 min	**First 30 min vs Mid 60 min**	**Mid 60 min vs last 60 min**
Block end choice probability P value : <0.001 All reversals tau P value : 0.521 Reward probability reversal tau P value: 0.228 Trans. probability reversal tau P value: 0.444	Block end choice probability P value : <0.001 All reversals tau P value : 0.709 Reward probability reversal tau P value: 0.705 Trans. probability reversal tau P value: 0.871	Block end choice probability P value : 0.013 All reversals tau P value : 0.811 Reward probability reversal tau P value: 0.433 Trans. probability reversal tau P value: 0.455

First 30 min, mid 60 min, last 60 min –Logistic regression model - within subject and cycles



First 30 min vs last 60 min	First 30 min vs Mid 60 min	Mid 60 min vs last 60 min
P values (1000 permutations): bias: 0.000 side: 0.027 correct: 0.000 choice : 0.000 outcome: 0.015 trans_CR : 0.000 trCR_x_out: 0.711	P values (1000 permutations): bias: 0.100 side: 0.151 correct: 0.000 choice : 0.832 outcome: 0.428 trans_CR : 0.000 trCR_x_out: 0.353	P values (1000 permutations): bias: 0.000 side: 0.434 correct: 0.825 choice : 0.000 outcome: 0.097 trans_CR : 0.962 trCR_x_out: 0.142

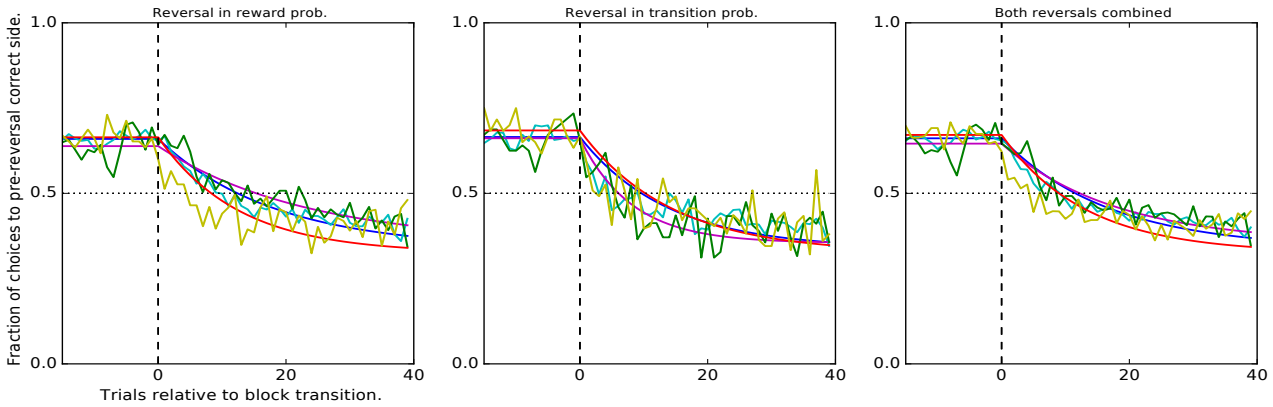
First 30 min, mid 60 min, last 60 min – RL model - within subject and cycles



First 30 min vs mid 60 min		Mid 60 min vs Last 60 min		First 30 min vs last 60 min	
P values (1000 permutations):		P values (1000 permutations):		P values (1000 permutations):	
alpQ	0.197	alpQ	0.247	**alpQ**	**0.008**
decQ	0.215	decQ	0.749	decQ	0.191
lbd	0.792	lbd	0.099	lbd	0.176
act	0.708	act	0.019	**act**	**0.045**
alpT	0.176	alpT	0.879	alpT	0.069
decT	0.713	decT	0.967	decT	0.561
alpCK	0.268	alpCK	0.083	alpCK	0.795
alpSK	0.102	alpSK	0.372	alpSK	0.025
G_td	0.205	G_td	0.131	**G_td**	**0.007**
G_mb	0.625	G_mb	0.488	G_mb	0.179
bs	0.601	bs	0.000	**bs**	**0.000**
sk	0.010	sk	0.789	**sk**	**0.010**
ec	0.286	ec	0.006	ec	0.348
sc	0.159	sc	0.735	sc	0.060
Model-based influence:	0.157	Model-based influence:	0.935	Model-based influence:	0.204
Model-free influence :	0.000	Model-free influence :	0.045	Model-free influence :	0.000

7.5.3. Injection stress study

7.5.3.1. Reversal plots

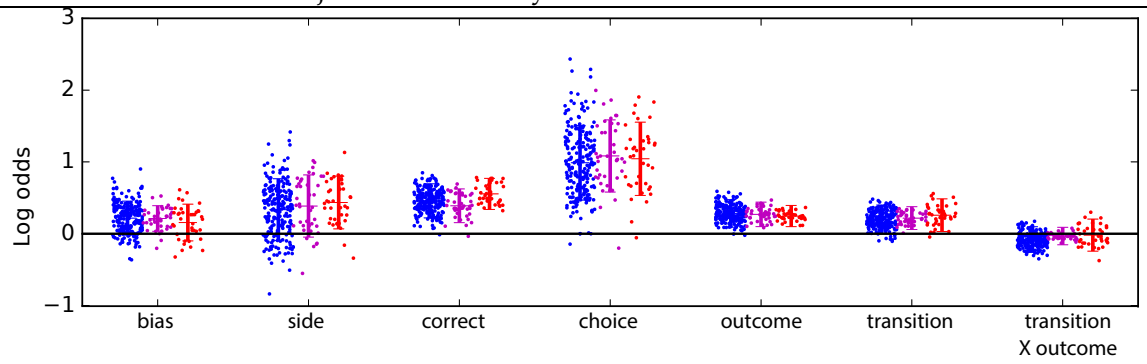


	None	Veh-90		Veh-15
Average block end prob.:	0.66	0.65		0.68
Reward prob. Reversal τ:	18.2	23.0	*→	12.6
Transition prob. Reversal τ:	13.7	8.1		14.2
Combined reversals τ:	16.8	17.7		13.2

Analysis by 1000 within subject and cycle permutations.	None versus Veh-15	None versus Veh-90	Veh-15 versus Veh-90
Block end choice probability P value:	0.442	0.308	0.156
All reversals tau P value:	0.195	0.721	0.166
Reward probability reversal tau P value:	0.086	0.134	0.016
Trans. probability reversal tau P value:	0.916	0.206	0.168

7.5.3.2. Logistic regression

No injection, injection 90 min before, injection 15 min before
Permutations within subject and within cycle.

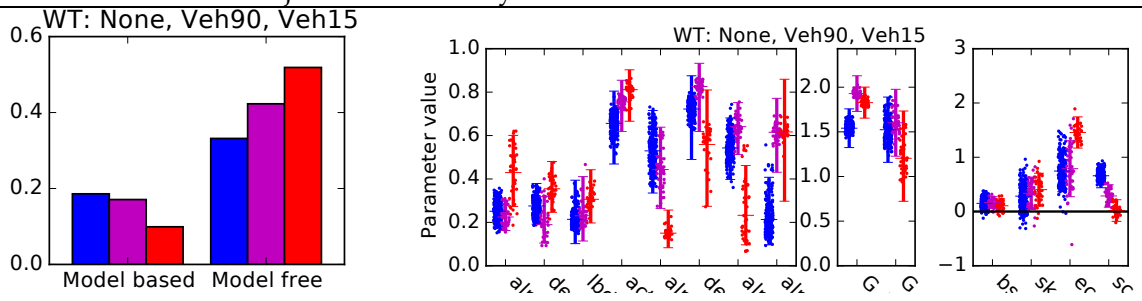


Acute (injection 15 min before) or recent (injection 90 min before) stress did not result in significant changes in logistic regression parameters

None vs Veh-15	None vs Veh-90	Veh-15 vs Veh-90
P values (1000 permutations):	P values (1000 permutations):	P values (1000 permutations):
bias : 0.150	bias : 0.907	bias : 0.306
side : 0.224	side : 0.573	side : 0.634
correct : 0.224	correct : 0.340	correct : 0.079
choice : 0.479	choice : 0.161	choice : 0.680
outcome : 0.531	outcome : 0.777	outcome : 0.724
trans_CR : 0.426	trans_CR : 0.779	trans_CR : 0.688
trCR_x_out: 0.363	trCR_x_out: 0.462	trCR_x_out: 0.895

7.5.4. Reinforcement learning model

No injection, injection 90 min before and injection 15 min before -
Permutations within subject and within cycle



None vs. Veh-90	None vs. Veh-15	Veh-90 vs Veh-15
P values (1000 permutations):	P values (1000 permutations):	P values (1000 permutations):
alpQ : 0.768	alpQ : 0.082	alpQ : 0.067
decQ : 0.254	decQ : 0.360	decQ : 0.347
lbd : 0.852	lbd : 0.327	lbd : 0.321
act : 0.371	act : 0.112	act : 0.142
alpT : 0.431	alpT : 0.056	alpT : 0.044
decT : 0.350	decT : 0.293	decT : 0.286
alpCK : 0.425	alpCK : 0.015	alpCK : 0.020
alpSK : 0.085	alpSK : 0.105	alpSK : 0.145
G_td : 0.109	G_td : 0.219	G_td : 0.222
G_mb : 0.862	G_mb : 0.482	G_mb : 0.430
bs : 0.618	bs : 0.097	bs : 0.085
sk : 0.412	sk : 0.054	sk : 0.059
ec : 0.912	ec : 0.015	ec : 0.016
sc : 0.341	sc : 0.054	sc : 0.068
Model-based influence: 0.662	Model-based influence: 0.047	Model-based influence: 0.040
Model-free influence : 0.121	Model-free influence : 0.005	Model-free influence : 0.006

7.6. Appendix to Chapter 6

7.6.1. COMT baseline

7.6.1.1. Cumulative trials

The two most active mice were both WTs and littermates[2]. They pulled the mean cumulative trials of the WT group up, compared to the Met group. When the two most active WT mice were removed the WT and Met had similar mean±SD of cumulative trials. This suggested that litter is a possible confounder to what could appear to be a genotype effect.

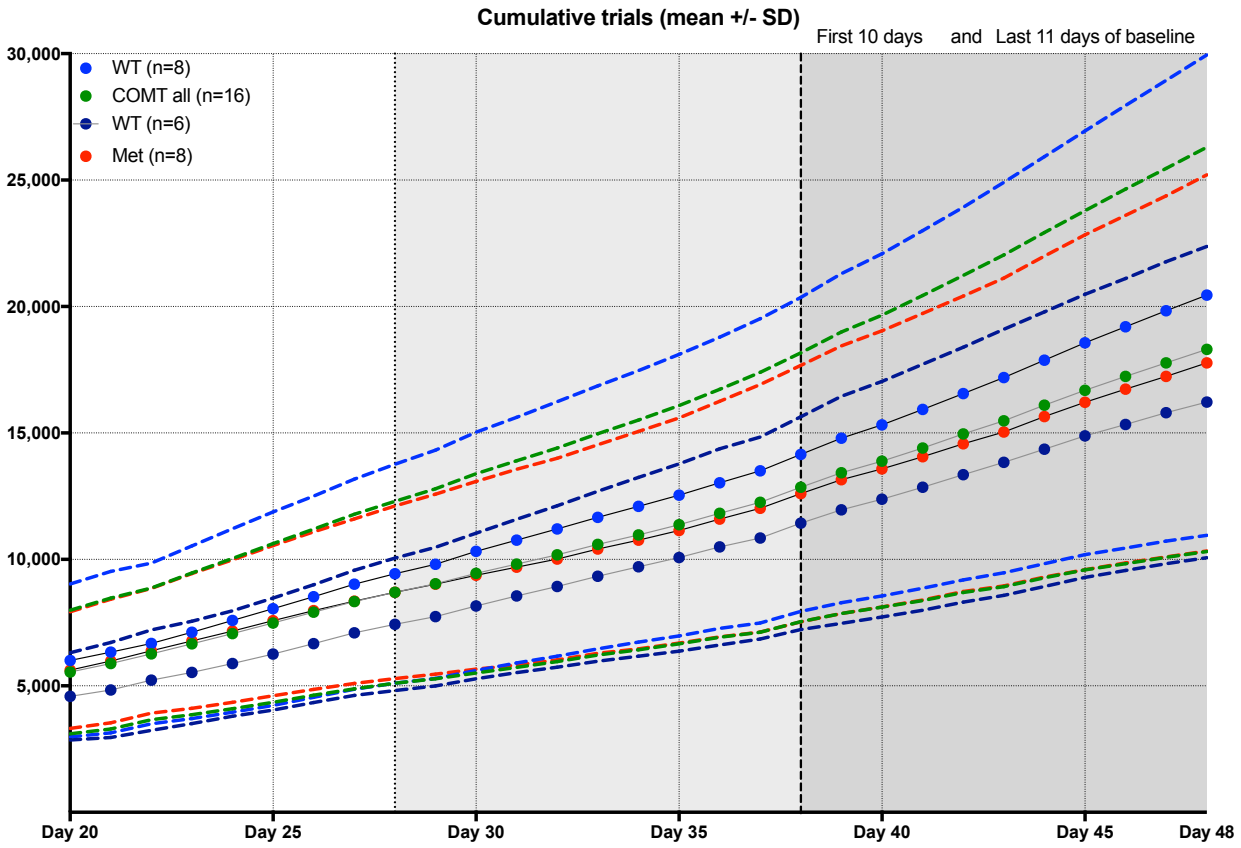
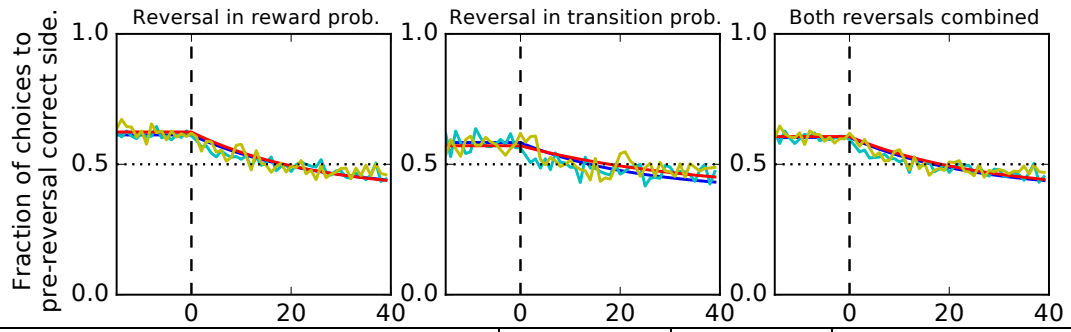


Figure 73: Cumulative trials during day 20-48. COMT-WT (blue), COMT-WT without the 2 most active mice (dark blue), COMT-Met (red), all mice (WT and Met: green). Mean (dots) ± SD (dashed).

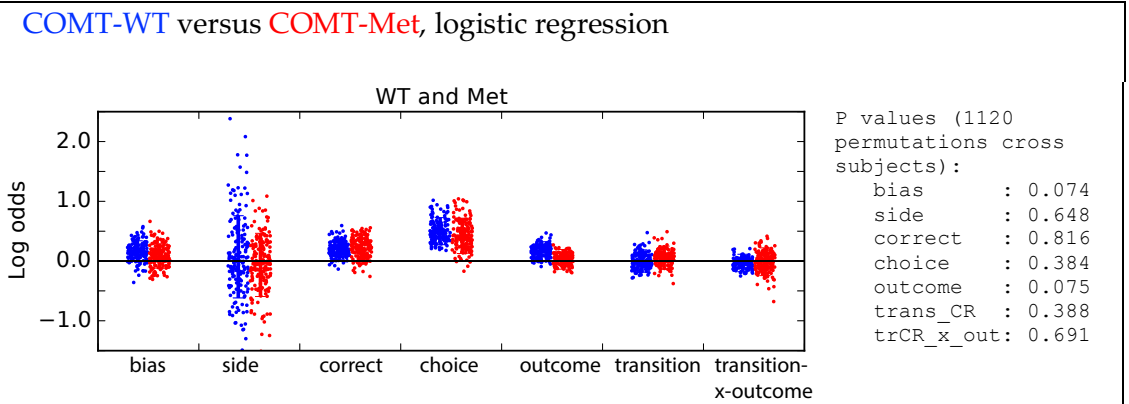
[2] There was no Met littermate.

7.6.1.2. Reversal analysis



*Cross subject permutations	WT	Met	*P=
Average block end choice probability:	0.604	0.608	0.792
Reward probability reversal, tau:	24.64	24.42	0.698
Trans. probability reversal, tau:	23.48	35.46	0.984
Combined reversals, tau:	24.30	26.95	0.152

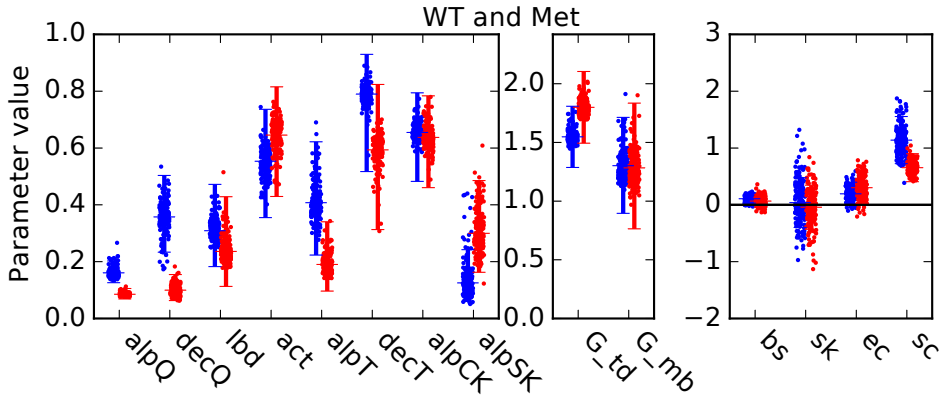
7.6.1.3. Logistic regression

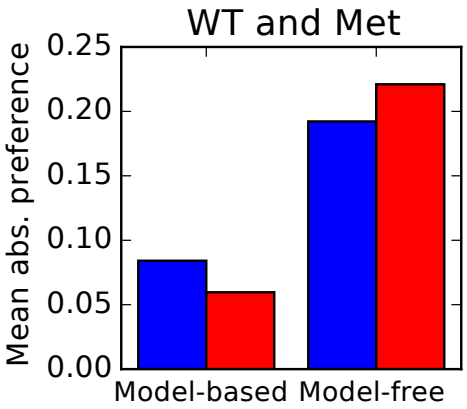


7.6.1.4. Reinforcement learning model

COMT-WT versus COMT-Met, RL model

Permutations cross subject






```
P values (1000 permutations):
alpQ                    : 0.497
decQ                    : 0.136
lbd                     : 0.329
act                     : 0.328
alpT                    : 0.030
decT                    : 0.101
alpCK                   : 0.906
alpSK                   : 0.488
G_td                    : 0.679
G_mb                    : 0.970
bs                      : 0.218
sk                      : 0.716
ec                      : 0.693
sc                      : 0.487
Model-based influence: 0.301
Model-free influence : 0.386
```

7.6.2. COMT stress

7.6.2.1. Circadian rhythm of plasma CORT

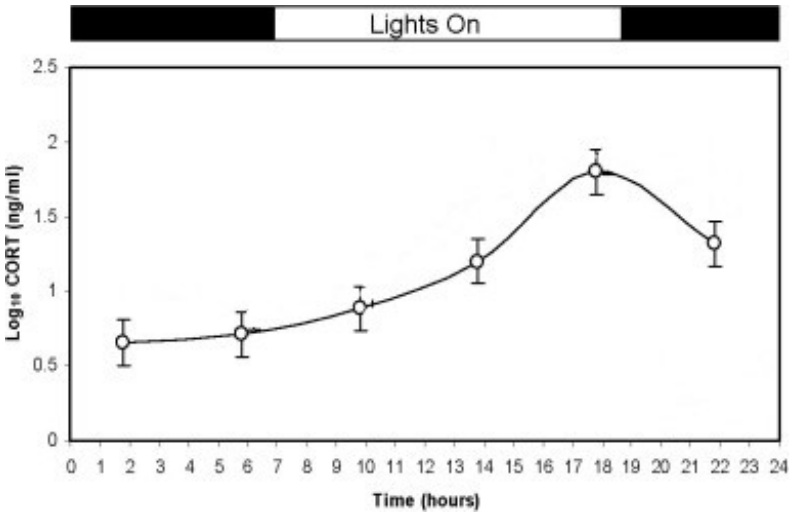
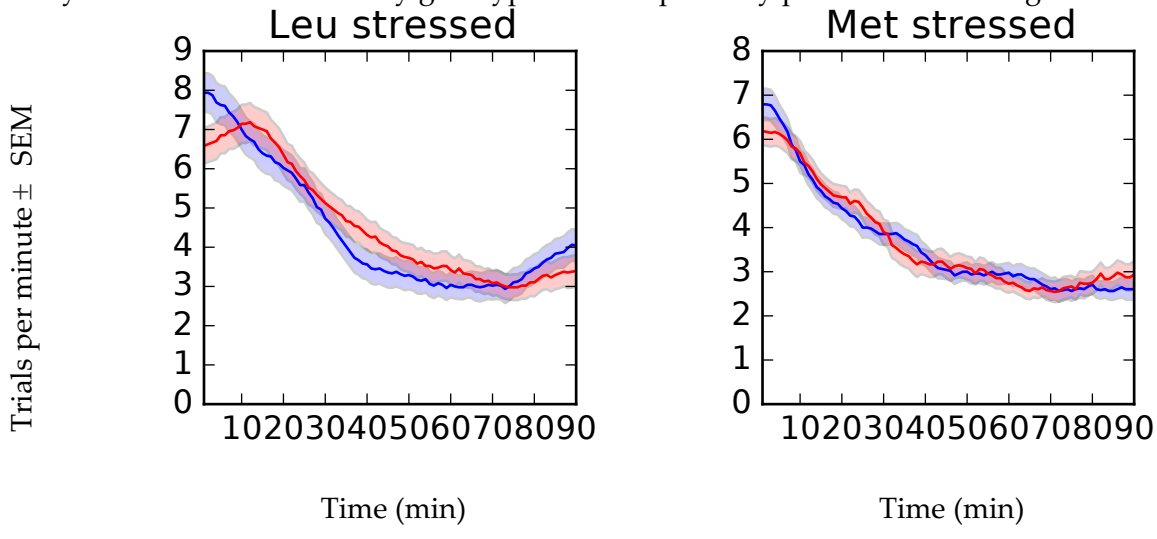


Figure 74: Circadian rhythm of plasma CORT in mice during a standard animal facilty 12-hour light off and on schedule. Adapted from (Malisch et al. 2008).

7.6.2.2. Trial rate divided by genotype

COMT: No injection, Injection 10 min before

Analysis of trial rate divided by genotype and compared by permutation-testing.



No injection vs injection (within-subject):
Trial number difference P value: 0.877

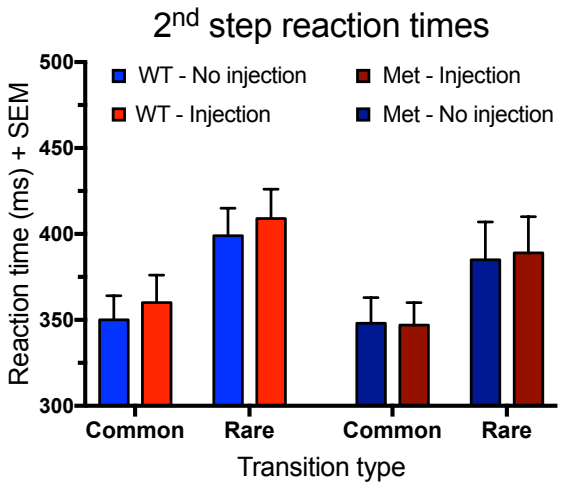
No injection vs injection (within-subject):
Trial number difference P value: 0.550

WT vs Met - No injection (across subject):
Trial number difference P value: 0.320

WT vs Met - Injection (across subject):
Trial number difference P value: 0.281

7.6.2.3. 2nd step Reaction time

COMT: No injection, Injection 10 min before
Analysis of 2nd step reaction time divided by genotype.




```
WT - No injection vs injection
(withinsubject):
Common RT P value:0.162
Rare   RT P value:0.513
Diff.  RT P value:0.878

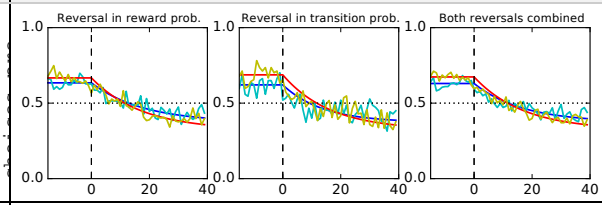
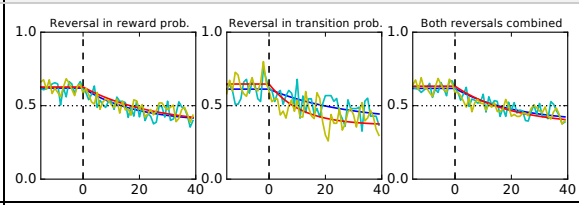
Met - No injection vs injection
(withinsubject):
Common RT P value:0.964
Rare   RT P value:0.770
Diff.  RT P value:0.734

WT vs Met - No injection (across subject):
Common RT P value:0.905
Rare   RT P value:0.641
Diff.  RT P value:0.477

WT vs Met - Injection (across subject):
Common RT P value:0.544
Rare   RT P value:0.491
Diff.  RT P value:0.713
```

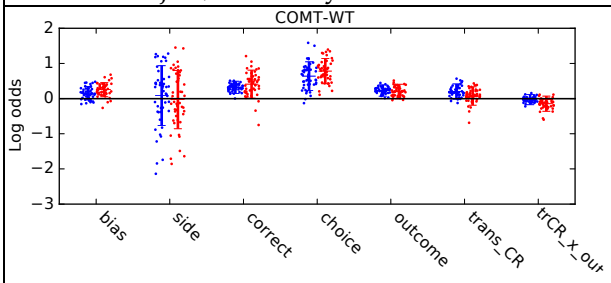
7.6.2.4. Reversal analysis divided by genotype

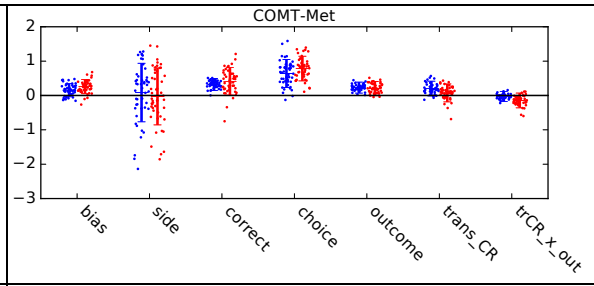
COMT: No injection, Injection 10 min before – Analysis of reversals divided by genotype.
Permutation test within subject and within cycle.

	COMT-WT (n=8)	COMT-Met (n=8)
Fraction of correct choice		
No inj. values	Average block end choice probability: 0.628 Reward probability reversal, tau: 17.906 Trans. probability reversal, tau: 13.812 Combined reversals, tau: 16.779	Average block end choice probability: 0.612 Reward probability reversal, tau: 18.510 Trans. probability reversal, tau: 28.015 Combined reversals, tau: 21.064
Injection values	Average block end choice probability: 0.679 Reward probability reversal, tau: 17.318 Trans. probability reversal, tau: 16.347 Combined reversals, tau: 17.009	Average block end choice probability: 0.634 Reward probability reversal, tau: 25.004 Trans. probability reversal, tau: 11.461 Combined reversals, tau: 20.459
Within subject	WT No injection vs Injection Block end choice probability P value :0.008 All reversals tau P value :0.935 Reward probability reversal tau P value:0.860 Trans. probability reversal tau P value:0.681	Met No injection vs Injection Block end choice probability P value :0.329 All reversals tau P value :0.903 Reward probability reversal tau P value:0.299 Trans. probability reversal tau P value:0.032
Cross subject	WT No Injection vs Met No Injection Block end choice probability P value : 0.570 All reversals tau P value : 0.459 Reward probability reversal tau P value: 0.925 Trans. probability reversal tau P value: 0.136	WT Injection vs Met Injection Block end choice probability P value : 0.083 All reversals tau P value : 0.367 Reward probability reversal tau P value: 0.106 Trans. probability reversal tau P value: 0.361

7.6.2.5. Logistic regression – divided by genotype

COMT: No injection, Injection 10 min before - Logistic regression model of first 90 min - within subject, within cycles





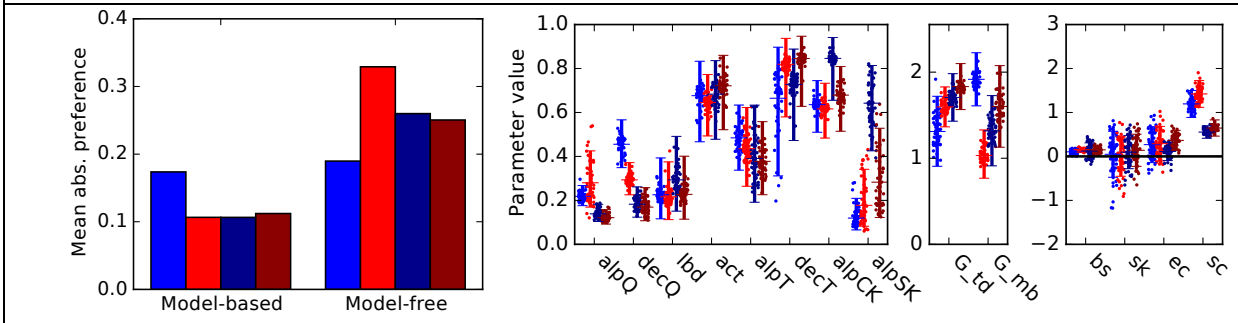

```
WT no injection vs Injection
P values (910 permutations):
   bias       : 0.036
   side       : 0.282
   correct    : 0.277
   choice     : 0.029
   outcome    : 0.587
   trans_CR   : 0.147
   trCR_x_out: 0.103
```

```
MET no injection vs Injection
P values (1000 permutations):
   bias       : 0.233
   side       : 0.405
   correct    : 0.943
   choice     : 0.994
   outcome    : 0.957
   trans_CR   : 0.010
   trCR_x_out: 0.514
```

```
WT no injection vs Met no Injection
P values (1000 permutations):
   bias       : 0.073
   side       : 0.705
   correct    : 0.895
   choice     : 0.405
   outcome    : 0.101
   trans_CR   : 0.485
   trCR_x_out: 0.683
```

7.6.2.6. RL model analysis - divided by genotype

COMT-WT: No injection, Injection; COMT-Met: No injection, Injection



WT: No injection vs Injection	Met: No injection vs Injection	No injection: WT vs Met	Injection: WT vs Met
P values (1000 permutations): alpQ : 0.496 decQ : 0.044 lbd : 0.965 act : 0.690 alpT : 0.636 decT : 0.364 alpCK: 0.799 alpSK: 0.282 G_td : 0.451 G_mb : 0.079 bs : 0.081 sk : 0.372 ec : 0.904 sc : 0.383 MB infl.: 0.099 MF infl.: 0.037	P values (1000 permutations): alpQ : 0.639 decQ : 0.592 lbd : 0.512 act : 0.439 alpT : 0.637 decT : 0.448 alpCK : 0.261 alpSK : 0.481 G_td : 0.497 G_mb : 0.372 bs : 0.279 sk : 0.447 ec : 0.079 sc : 0.674 MB infl.: 0.789 MF infl.: 0.810	P values (1000 permutations): alpQ : 0.335 decQ : 0.141 lbd : 0.501 act : 0.808 alpT : 0.429 decT : 0.728 alpCK : 0.356 alpSK : 0.047 G_td : 0.377 G_mb : 0.229 bs : 0.673 sk : 0.843 ec : 0.612 sc : 0.176 MB infl.: 0.078 MF infl.: 0.329	P values (1000 permutations): alpQ : 0.302 decQ : 0.093 lbd : 0.957 act : 0.255 alpT : 0.525 decT : 0.832 alpCK : 0.794 alpSK : 0.333 G_td : 0.739 G_mb : 0.213 bs : 0.958 sk : 0.471 ec : 0.738 sc : 0.081 MB infl.: 0.791 MF infl.: 0.340

Bibliography

Abercrombie, E. D., et al. (1989), 'Differential effect of stress on in vivo dopamine release in striatum, nucleus accumbens, and medial frontal cortex', *J Neurochem,* 52 (5), 1655-8.

Adams, C. D. and Dickinson, A. (1981), 'Actions and habits: Variations in associative representations during instrumental learning', *Information processing in animals: Memory mechanisms*, 143-65.

Ahn, Soyon and Phillips, Anthony G (1999), 'Dopaminergic correlates of sensory-specific satiety in the medial prefrontal cortex and nucleus accumbens of the rat', *Journal of Neuroscience,* 19, RC29 (1-6).

Akam, T. (2016a), 'pyControl code', (https://bitbucket.org/takam/pycontrol).

Akam, T (2016b), 'Two-step task analysis code', (https://bitbucket.org/takam/two-step-task-analysis).

Akam T., Vas I., Zhang X., Pereira M., Oliveira R., Dayan P., Costa R. (2016), 'Anterior Cingulate Silencing Disrupts Modelbased Reinforcement Learning in a Two-step Decision Task'.

Akam, Thomas, Costa, Rui, and Dayan, Peter (2015), 'Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task', *PLoS computational biology,* 11 (12).

Al-Hendy, A. and Salama, S. A. (2006), 'Catechol-O-methyltransferase polymorphism is associated with increased uterine leiomyoma risk in different ethnic groups', *J Soc Gynecol Investig,* 13 (2), 136-44.

Anderson, Marti J. (2001), 'Permutation tests for univariate or multivariate analysis of variance and regression', *Canadian Journal of Fisheries and Aquatic Sciences,* 58 (3), 626-39.

Apud, J. A., et al. (2007), 'Tolcapone improves cognition and cortical information processing in normal human subjects', *Neuropsychopharmacology,* 32 (5), 1011-20.

Armbruster, Diana, et al. (2011), 'Children under stress - COMT genotype and stressful life events predict cortisol increase in an acute social stress paradigm', *The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP),* 15 (9), 1229-39.

Arnsten, A. F. (2009a), 'Stress signalling pathways that impair prefrontal cortex structure and function', *Nature reviews. Neuroscience,* 10 (6), 410-22.

Arnsten, A. F. and Goldman-Rakic, P. S. (1998), 'Noise stress impairs prefrontal cortical cognitive function in monkeys: evidence for a hyperdopaminergic mechanism', *Arch Gen Psychiatry,* 55 (4), 362-8.

Arnsten, A. F. T. (2015), 'Stress weakens prefrontal networks: molecular insults to higher cognition', *Nature Neuroscience,* 18 (10), 1376-85.

Arnsten, Amy F. T. (2009b), 'Stress signalling pathways that impair prefrontal cortex structure and function', *Nature reviews. Neuroscience,* 10 (6), 410-22.

Axelrod, J. (1957), 'O-Methylation of Epinephrine and Other Catechols in vitro and in vivo', *Science,* 126 (3270), 400-01.

Axelrod, J. and Tomchick, R. (1958), 'Enzymatic O-methylation of epinephrine and other catechols', *J Biol Chem,* 233 (3), 702-5.

Ayars, Alisabeth (2016), 'Can model-free reinforcement learning explain deontological moral judgments?', *Cognition,* 150, 232-42.

Balcombe, Jonathan P., Barnard, Neal D., and Sandusky, Chad (2004), 'Laboratory routines cause animal stress', *Contemporary topics in laboratory animal science / American Association for Laboratory Animal Science,* 43 (6), 42-51.

Balleine, B. (1992), 'Instrumental performance following a shift in primary motivation depends on

incentive learning', *Journal of experimental psychology. Animal behavior processes,* 18 (3), 236-50.
Balleine, B. W. and Dickinson, A. (1998), 'The role of incentive learning in instrumental outcome revaluation by sensory-specific satiety', *Animal Learning & Behavior,* 26 (1), 46-59.
Balleine, B. W. and O'Doherty, J. P. (2010), 'Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action', *Neuropsychopharmacology,* 35 (1), 48-69.
Barkus, C., et al. (2016), 'Genotype-Dependent Effects of COMT Inhibition on Cognitive Function in a Highly Specific, Novel Mouse Model of Altered COMT Activity', *Neuropsychopharmacology*.
Barnett, J. H., Scoriels, L., and Munafo, M. R. (2008), 'Meta-analysis of the cognitive effects of the catechol-O-methyltransferase gene Val158/108Met polymorphism', *Biol Psychiatry,* 64 (2), 137-44.
Barnett, J. H., et al. (2007), 'Effects of the catechol-O-methyltransferase Val158Met polymorphism on executive function: a meta-analysis of the Wisconsin Card Sort Test in schizophrenia and healthy controls', *Mol Psychiatry,* 12 (5), 502-9.
Bassareo, Valentina and Di Chiara, Gaetano (1999), 'Modulation of feeding-induced activation of mesolimbic dopamine transmission by appetitive stimuli and its relation to motivational state', *European Journal of Neuroscience,* 11 (12), 4389-97.
Björklund, A. and Dunnett, S. B. (2007), 'Dopamine neuron systems in the brain: an update', *Trends in Neurosciences,* 30 (2), 194-202.
Braun, Stephanie and Hauber, Wolfgang (2013), 'Acute stressor effects on goal-directed action in rats'.
Brozoski, T. J., et al. (1979), 'Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey', *Science,* 205 (4409), 929-32.
Buckert, M., et al. (2012), 'The COMT Val158Met polymorphism modulates working memory performance under acute stress', *Psychoneuroendocrinology,* 37 (11), 1810-21.
Budygin, E. A., et al. (1999a), '[The effect of the dopamine neuronal reuptake inhibitor GBR 12909 on the dopaminergic neurotransmission parameters in the rat striatum in vivo]', *Eksp Klin Farmakol,* 62 (5), 3-6.
Budygin, E. A., et al. (1999b), 'Effect of tolcapone, a catechol-O-methyltransferase inhibitor, on striatal dopaminergic transmission during blockade of dopamine uptake', *Eur J Pharmacol,* 370 (2), 125-31.
Butts, K. A. and Phillips, A. G. (2013), 'Glucocorticoid receptors in the prefrontal cortex regulate dopamine efflux to stress via descending glutamatergic feedback to the ventral tegmental area', *Int J Neuropsychopharmacol,* 16 (8), 1799-807.
Chattarji, S., et al. (2015), 'Neighborhood matters: divergent patterns of stress-induced plasticity across the brain', *Nat Neurosci,* 18 (10), 1364-75.
Chen, Chunhui, et al. (2011a), 'Sex Modulates the Associations Between the COMT Gene and Personality Traits', *Neuropsychopharmacology,* 36 (8), 1593-98.
Chen, J., et al. (2011b), 'Orientation and cellular distribution of membrane-bound catechol-O-methyltransferase in cortical neurons: implications for drug development', *J Biol Chem,* 286 (40), 34752-60.
Chen, J., et al. (2004), 'Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain', *Am J Hum Genet,* 75 (5), 807-21.
Colwill, Ruth M. and Rescorla, Robert A. (1985), 'Postconditioning devaluation of a reinforcer affects instrumental responding', *Journal of Experimental Psychology: Animal Behavior Processes,* 11 (1), 120.
Corral-Frias, N. S., et al. (2016), 'COMT Val(158) Met genotype is associated with reward learning:

a replication study and meta-analysis', *Genes Brain Behav,* 15 (5), 503-13.
Crawley, J.N. (2007), *What's Wrong With My Mouse?: Behavioral Phenotyping of Transgenic and Knockout Mice* (Wiley).
Culbreth, Adam J., et al. (2016), 'Reduced model-based decision-making in schizophrenia', *Journal of Abnormal Psychology,* 125 (6), 777-87.
Daw, N. D., Niv, Y., and Dayan, P. (2005), 'Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control', *Nat Neurosci,* 8 (12), 1704-11.
Daw, N. D., et al. (2011), 'Model-Based Influences on Humans' Choices and Striatal Prediction Errors', *Neuron,* 69 (6), 1204-15.
Dayan, Peter and Berridge, Kent C. (2014), 'Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation', *Cognitive, Affective, & Behavioral Neuroscience,* 14 (2), 473-92.
de Araujo, Ivan E, et al. (2012), 'The gut–brain dopamine axis: a regulatory system for caloric intake', *Physiology & behavior,* 106 (3), 394-99.
Desbonnet, L., et al. (2012), 'Physiological and behavioural responsivity to stress and anxiogenic stimuli in COMT-deficient mice', *Behav Brain Res,* 228 (2), 351-8.
Deserno, L., et al. (2015a), 'Lateral prefrontal model-based signatures are reduced in healthy individuals with high trait impulsivity', *Transl Psychiatry,* 5, e659.
Deserno, L., et al. (2015b), 'Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making', *Proc Natl Acad Sci U S A,* 112 (5), 1595-600.
Dias-Ferreira, E., et al. (2009), 'Chronic stress causes frontostriatal reorganization and affects decision-making', *Science,* 325 (5940), 621-5.
Doll, B. B., et al. (2016), 'Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning', *Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning,* 36 (4).
Drude, Solveig, et al. (2011), 'Side effects of control treatment can conceal experimental data when studying stress responses to injection and psychological stress in mice', *Lab animal,* 40 (4), 119-28.
Egan, Michael F., et al. (2001), 'Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia', *Proceedings of the National Academy of Sciences,* 98 (12), 6917-22.
Eppinger, Ben, et al. (2013), 'Of goals and habits: age-related and individual differences in goal-directed decision-making', *Frontiers in Neuroscience,* 7 (253).
Farrell, S. M., et al. (2012), 'COMT Val(158)Met genotype determines the direction of cognitive effects produced by catechol-O-methyltransferase inhibition', *Biol Psychiatry,* 71 (6), 538-44.
Floderus, Y., Ross, S. B., and Wetterberg, L. (1981), 'Erythrocyte catechol-O-methyltransferase activity in a Swedish population', *Clin Genet,* 19 (5), 389-92.
Frank, Michael J., et al. (2007), 'Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning', *Proceedings of the National Academy of Sciences of the United States of America,* 104 (41), 16311-16.
G E Alexander, M R DeLong, and, and Strick, P L (1986), 'Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex', *Annual Review of Neuroscience,* 9 (1), 357-81.
Gillan, Claire M., et al. (2015), 'Model-based learning protects against forming habits', *Cognitive, Affective, & Behavioral Neuroscience,* 15 (3), 523-36.
Gillan, Claire M., et al. (2016), 'Characterizing a psychiatric symptom dimension related to deficits in goal-directed control', *eLife,* 5.
Glimcher, P. W. (2011), 'Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis', *Proc Natl Acad Sci U S A,* 108 Suppl 3, 15647-54.

Gogos, J. A., et al. (1998), 'Catechol-O-methyltransferase-deficient mice exhibit sexually dimorphic changes in catecholamine levels and behavior', *Proc Natl Acad Sci U S A,* 95 (17), 9991-6.

Goldman-Rakic, P. S. (1998), 'The cortical dopamine system: role in memory and cognition', *Adv Pharmacol,* 42, 707-11.

Goldman-Rakic, P. S., Muly, E. C., 3rd, and Williams, G. V. (2000), 'D(1) receptors in prefrontal cells and circuits', *Brain Res Brain Res Rev,* 31 (2-3), 295-301.

Gourley, Shannon L., Swanson, Andrew M., and Koleske, Anthony J. (2013), 'Corticosteroid-Induced Neural Remodeling Predicts Behavioral Vulnerability and Resilience', *The Journal of Neuroscience,* 33 (7), 3107-12.

Grossman, M. H., Emanuel, B. S., and Budarf, M. L. (1992), 'Chromosomal mapping of the human catechol-O-methyltransferase gene to 22q11.1----q11.2', *Genomics,* 12 (4), 822-5.

Gurvich, C. and Rossell, S. L. (2015), 'Dopamine and cognitive control: sex-by-genotype interactions influence the capacity to switch attention', *Behavioural brain research,* 281, 96-101.

Haber, Suzanne N. and Calzavara, Roberta (2009), 'The cortico-basal ganglia integrative network: The role of the thalamus', *Brain Research Bulletin,* 78 (2–3), 69-74.

Harrison, Paul J. and Tunbridge, Elizabeth M. (2008), 'Catechol-O-methyltransferase (COMT): a gene contributing to sex differences in brain function, and to sexual dimorphism in the predisposition to psychiatric disorders', *Neuropsychopharmacology,* 33 (13), 3037-45.

Hattori, T. (1993), 'Conceptual history of the nigrostriatal dopamine system', *Neurosci Res,* 16 (4), 239-62.

Hawkins, Douglas M. (2004), 'The Problem of Overfitting', *Journal of Chemical Information and Computer Sciences,* 44 (1), 1-12.

Herman, J. P., et al. (2003), 'Central mechanisms of stress integration: hierarchical circuitry controlling hypothalamo-pituitary-adrenocortical responsiveness', *Front Neuroendocrinol,* 24 (3), 151-80.

Hernaus, D., et al. (2013), 'COMT Val158Met genotype selectively alters prefrontal [18F]fallypride displacement and subjective feelings of stress in response to a psychosocial stress challenge', *PLoS One,* 8 (6), e65662.

Hersch, S. M., et al. (1997), 'Subcellular localization and molecular topology of the dopamine transporter in the striatum and substantia nigra', *J Comp Neurol,* 388 (2), 211-27.

Hiroi, N., et al. (2013), 'Copy number variation at 22q11.2: from rare variants to common mechanisms of developmental neuropsychiatric disorders', *Molecular psychiatry,* 18 (11), 10.1038/mp.2013.92.

Hong, J., et al. (1998), 'Distribution of catechol-O-methyltransferase expression in human central nervous system', *Neuroreport,* 9 (12), 2861-4.

Huber, Anna (2015), 'Effects of Catechol-O-Methyltransferase on Reinforcement Learning and Mesolimbic Dopamine Transimission', (University of Oxford).

Hunter, J. D. (2007), 'Matplotlib: A 2D graphics environment', *Computing In Science & Engineering,* 9 (3), 90-95.

Huys, Q. J., et al. (2012), 'Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees', *PLoS Comput Biol,* 8 (3), e1002410.

Inglis, F. M. and Moghaddam, B. (1999), 'Dopaminergic innervation of the amygdala is highly responsive to stress', *J Neurochem,* 72 (3), 1088-94.

Ito, M. and Doya, K. (2009), 'Validation of decision-making models and analysis of decision variables in the rat basal ganglia', *J Neurosci,* 29 (31), 9861-74.

Jacobs, Emily and D'Esposito, Mark (2011), 'Estrogen Shapes Dopamine-Dependent Cognitive Processes: Implications for Women's Health', *The Journal of Neuroscience,* 31 (14), 5286-93.

Kaakkola, S., Gordin, A., and Mannisto, P. T. (1994), 'General properties and clinical possibilities of

new selective inhibitors of catechol O-methyltransferase', *Gen Pharmacol,* 25 (5), 813-24.
Käenmäki, Mikko, et al. (2010), 'Quantitative role of COMT in dopamine clearance in the prefrontal cortex of freely moving mice', *Journal of Neurochemistry,* 114 (6), 1745-55.
Karoum, Farouk, Chrapusta, Stanislaw J., and Egan, Michael F. (1994), '3-Methoxytyramine Is the Major Metabolite of Released Dopamine in the Rat Frontal Cortex: Reassessment of the Effects of Antipsychotics on the Dynamics of Dopamine Release and Metabolism in the Frontal Cortex, Nucleus Accumbens, and Striatum by a Simple Two Pool Model', *Journal of Neurochemistry,* 63 (3), 972-79.
Kastner, A., et al. (1994), 'Immunohistochemical study of catechol-O-methyltransferase in the human mesostriatal system', *Neuroscience,* 62 (2), 449-57.
Keramati, M., Dezfouli, A., and Piray, P. (2011), 'Speed/accuracy trade-off between the habitual and the goal-directed processes', *PLoS Comput Biol,* 7 (5), e1002055.
Khromova, Irina, et al. (1997), 'Effects of selective catechol-O-methyltransferase inhibitors on single-trial passive avoidance retention in male rats', *Behavioural Brain Research,* 86 (1), 49-57.
Kirschbaum, C., Pirke, K. M., and Hellhammer, D. H. (1993), 'The 'Trier Social Stress Test'--a tool for investigating psychobiological stress responses in a laboratory setting', *Neuropsychobiology,* 28 (1-2), 76-81.
Krugel, Lea K., et al. (2009), 'Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions', *Proceedings of the National Academy of Sciences,* 106 (42), 17951-56.
Laatikainen, L. M. (2010), 'The role of catechol-O-methyltransferase (COMT) in hippocampal function', (University of Oxford).
Laatikainen, Linda M., et al. (2013), 'Sexually Dimorphic Effects of Catechol-O-Methyltransferase (COMT) Inhibition on Dopamine Metabolism in Multiple Brain Regions', *PLoS ONE,* 8 (4).
Lamb, Yvette Nicole, et al. (2016), 'Catechol-O-methyltransferase val158met polymorphism interacts with sex to affect face recognition ability', *Frontiers in Psychology,* 7.
Lapish, C. C., et al. (2009), 'Tolcapone enhances food-evoked dopamine efflux and executive memory processes mediated by the rat prefrontal cortex', *Psychopharmacology (Berl),* 202 (1-3), 521-30.
Legendre, P. and Legendre, L. (1998), *Numerical ecology* (2nd edn., Elsevier Science: Elsevier Science).
Lewis, D. A., et al. (2001), 'Dopamine transporter immunoreactivity in monkey cerebral cortex: regional, laminar, and ultrastructural localization', *J Comp Neurol,* 432 (1), 119-36.
Lindenberger, Ulman, et al. (2008), 'Age-related decline in brain resources modulates genetic effects on cognitive functioning', *Frontiers in Neuroscience,* 2 (39).
Lotta, T., et al. (1995), 'Kinetics of human soluble and membrane-bound catechol O-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme', *Biochemistry,* 34 (13), 4202-10.
Malisch, Jessica L., et al. (2008), 'Circadian pattern of total and free corticosterone concentrations, corticosteroid-binding globulin, and physical activity in mice selectively bred for high voluntary wheel-running behavior', *General and Comparative Endocrinology,* 156 (2), 210-17.
Manly, Bryan F. J. (2007a), *Randomization, bootstrap, and Monte Carlo methods in biology* (3rd edn., Texts in statistical science; Boca Raton, FL: Chapman & Hall/ CRC) 455 p.
--- (2007b), 'Randomization Tests, 4th Edition by Eugene S. Edgington, Patrick Onghena', *International Statistical Review,* 75 (2), 269-69.
Mannisto, P. T., et al. (1992), 'Characteristics of catechol O-methyl-transferase (COMT) and properties of selective COMT inhibitors', *Prog Drug Res,* 39, 291-350.
Matsumoto, M., et al. (2003), 'Catechol O-methyltransferase (COMT) mRNA expression in the

dorsolateral prefrontal cortex of patients with schizophrenia', *Neuropsychopharmacology,* 28 (8), 1521-30.

Mattay, V. S., et al. (2003), 'Catechol O-methyltransferase val158-met genotype and individual variation in the brain response to amphetamine', *Proc Natl Acad Sci U S A,* 100 (10), 6186-91.

McCutcheon, James E, Beeler, Jeff A, and Roitman, Mitchell F (2012), 'Sucrose-predictive cues evoke greater phasic dopamine release than saccharin-predictive cues', *Synapse,* 66 (4), 346-51.

Meijer, M. K., et al. (2006), 'Effect of restraint and injection methods on heart rate and body temperature in mice', *Laboratory animals,* 40 (4), 382-91.

Mier, D., Kirsch, P., and Meyer-Lindenberg, A. (2010), 'Neural substrates of pleiotropic action of genetic variation in COMT: a meta-analysis', *Mol Psychiatry,* 15 (9), 918-27.

Mueller, A., et al. (2012), 'Genetic contributions to acute autonomic stress responsiveness in children', *Int J Psychophysiol,* 83 (3), 302-8.

Murphy, B. L., et al. (1996a), 'Increased dopamine turnover in the prefrontal cortex impairs spatial working memory performance in rats and monkeys', *Proc Natl Acad Sci U S A,* 93 (3), 1325-9.

Murphy, B. L., et al. (1996b), 'Dopamine and spatial working memory in rats and monkeys: pharmacological reversal of stress-induced impairment', *J Neurosci,* 16 (23), 7768-75.

Murphy, Tytus, Dias, Gisele, and Thuret, Sandrine (2014), 'Effects of Diet on Brain Plasticity in Animal and Human Studies: Mind the Gap', *Neural Plasticity,* 2014, 1-32.

Myohanen, T. T. and Mannisto, P. T. (2010), 'Distribution and functions of catechol-O-methyltransferase proteins: do recent findings change the picture?', *Int Rev Neurobiol,* 95, 29-47.

Nagel, I. E., et al. (2008), 'Human aging magnifies genetic effects on executive functioning and working memory', *Front Hum Neurosci,* 2, 1.

O'Doherty, J., et al. (2004), 'Dissociable roles of ventral and dorsal striatum in instrumental conditioning', *Science,* 304 (5669), 452-4.

O'Doherty, J. P., Cockburn, J., and Wolfgang, M. P. (2017), 'Learning, Reward, and Decision Making', *Annual Review of Psychology,* 68 (19).

Olver, J. S., et al. (2015), 'Impairments of spatial working memory and attention following acute psychosocial stress', *Stress Health,* 31 (2), 115-23.

Otto, A. R., et al. (2013a), 'The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive', *Psychol Sci,* 24 (5), 751-61.

Otto, A. R., et al. (2013b), 'Working-memory capacity protects model-based learning from stress', *Proc Natl Acad Sci U S A,* 110 (52), 20941-6.

Otto, Ross A., et al. (2013c), 'The Curse of Planning Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive', *Psychological Science,* 24 (5), 751-61.

Otto, Ross A., et al. (2013d), 'Working-memory capacity protects model-based learning from stress', *Proceedings of the National Academy of Sciences of the United States of America,* 110 (52), 20941-46.

Palmatier, M. A., Kang, A. M., and Kidd, K. K. (1999), 'Global variation in the frequencies of functionally different catechol-O-methyltransferase alleles', *Biol Psychiatry,* 46 (4), 557-67.

Papageorgiou, Georgios K, et al. (2016), 'Mesolimbic Dopamine Encodes Prediction Errors in a State-Dependent Manner', *Cell Reports,* 15 (2), 221-28.

Papaleo, F., et al. (2012), 'Effects of sex and COMT genotype on environmentally modulated cognitive control in mice', *Proceedings of the National Academy of Sciences of the United States of America,* 109 (49), 20160-65.

Papaleo, F., et al. (2008), 'Genetic dissection of the role of catechol-O-methyltransferase in cognition and stress reactivity in mice', *J Neurosci,* 28 (35), 8709-23.

Papaleo, Francesco, et al. (2015), 'Sex-dichotomous effects of functional COMT genetic variations on cognitive functions disappear after menopause in both health and schizophrenia', *European Neuropsychopharmacology,* 25 (12), 2349-63.

Phillips, A. G., Ahn, S., and Floresco, S. B. (2004), 'Magnitude of dopamine release in medial prefrontal cortex predicts accuracy of memory on a delayed response task', *J Neurosci,* 24 (2), 547-53.

Phipson, Belinda and Smyth, Gordon K. (2010), 'Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn', *Statistical Applications in Genetics and Molecular Biology,* 9 (1).

Pizzagalli, D. A. (2014), 'Depression, stress, and anhedonia: toward a synthesis and integrated model', *Annu Rev Clin Psychol,* 10, 393-423.

Pochon, Jean-Baptiste, et al. (2001), 'The Role of Dorsolateral Prefrontal Cortex in the Preparation of Forthcoming Actions: an fMRI Study', *Cerebral Cortex,* 11 (3), 260-66.

Qin, S., et al. (2009), 'Acute psychological stress reduces working memory-related activity in the dorsolateral prefrontal cortex', *Biol Psychiatry,* 66 (1), 25-32.

Radenbach, C., et al. (2015), 'The interaction of acute and chronic stress impairs model-based behavioral control', *Psychoneuroendocrinology,* 53, 268-80.

Radua, J., et al. (2014), 'COMT Val158Met x SLC6A4 5-HTTLPR interaction impacts on gray matter volume of regions supporting emotion processing', *Soc Cogn Affect Neurosci,* 9 (8), 1232-8.

Ryabinin, Andrey E., Wang, Yuan-Mei, and Finn, Deborah A. (1999), 'Different Levels of Fos Immunoreactivity After Repeated Handling and Injection Stress in Two Inbred Strains of Mice', *Pharmacology Biochemistry and Behavior,* 63 (1), 143-51.

Sapolsky, Robert M. (2015), 'Stress and the brain: individual variability and the inverted-U', *Nature Neuroscience,* 18 (10), 1344-46.

Sawaguchi, T. and Goldman-Rakic, P. S. (1991), 'D1 dopamine receptors in prefrontal cortex: involvement in working memory', *Science,* 251 (4996), 947-50.

Schultz, W., Dayan, P., and Montague, P. R. (1997), 'A neural substrate of prediction and reward', *Science,* 275 (5306), 1593-9.

Schultz, Wolfram (2006), 'Behavioral theories and the neurophysiology of reward', *Annual review of psychology,* 57, 87-115.

Schwabe, L. and Wolf, O. T. (2009), 'Stress prompts habit behavior in humans', *J Neurosci,* 29 (22), 7191-8.

--- (2013), 'Stress and multiple memory systems: from 'thinking' to 'doing'', *Trends Cogn Sci,* 17 (2), 60-8.

Schwabe, Lars and Wolf, Oliver T. (2011), 'Stress-induced modulation of instrumental behavior: From goal-directed to habitual control of action', *Behavioural Brain Research,* 219 (2), 321-28.

Schwarz, G. (1978), 'Estimating Dimension of a Model', *Annals of Statistics,* 6 (2), 461-64.

Sealfon, S. C. and Olanow, C. W. (2000), 'Dopamine receptors: from structure to behavior', *Trends Neurosci,* 23 (10 Suppl), S34-40.

Sebold, M., et al. (2014), 'Model-Based and Model-Free Decisions in Alcohol Dependence', *Neuropsychobiology,* 70 (2), 122-31.

Sesack, S. R., et al. (1998), 'Dopamine axon varicosities in the prelimbic division of the rat prefrontal cortex exhibit sparse immunoreactivity for the dopamine transporter', *J Neurosci,* 18 (7), 2697-708.

Sharp, M. E., et al. (2015), 'Dopamine selectively remediates ‘model-based’ reward learning: a computational approach', *Brain,* 139 (2), 355-64.

Sharp, M. E., et al. (2016), 'Dopamine selectively remediates 'model-based' reward learning: a computational approach', *Brain,* 139 (Pt 2), 355-64.

Simpson, E. H., et al. (2014), 'Genetic variation in COMT activity impacts learning and dopamine

release capacity in the striatum', *Learning & Memory,* 21 (4).
Skatova, Anya, Chan, Patricia, and Daw, Nathaniel (2013), 'Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task', *Frontiers in Human Neuroscience,* 7 (525).
Smittenaar, Peter, et al. (2013), 'Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans', *Neuron,* 80 (4), 914-19.
Staal, M. A. (2004), 'Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework', (NASA Ames Research Center).
Stokes, A. F. and Kite, K. (2001), 'On grasping a nettle and becoming emotional', in Peter A. Hancock and Paula A. Desmond (eds.), *Stress, workload, and fatigue* (Mahwah, N.J.: Lawrence Erlbaum Associates).
Sutton, Richard S. and Barto, Andrew G. (1998), *Reinforcement learning : an introduction* (Cambridge, Mass. ; London: MIT Press) xviii, 322 p.
Tabbara, Rayane I., et al. (2016), 'Effects of sucrose concentration and water deprivation on Pavlovian conditioning and responding for conditioned reinforcement', *Behavioral neuroscience,* 130 (2), 231-42.
Tammimäki, Anne and Männistö, Pekka T. (2011), 'Effect of Genetic Modifications in the Synaptic Dopamine Clearance Systems on Addiction-like Behaviour in Mice', *Basic & Clinical Pharmacology & Toxicology,* 108 (1), 2-8.
Tenhunen, J., et al. (1994), 'Genomic organization of the human catechol O-methyltransferase gene and its expression from two distinct promoters', *Eur J Biochem,* 223 (3), 1049-59.
Tepas, D. I. and Price, J. M. (2001), 'What is stress and what is fatigue?', in Peter A. Hancock and Paula A. Desmond (eds.), *Stress, workload, and fatigue* (Mahwah, N.J.: Lawrence Erlbaum Associates).
Thompson, John Md, et al. (2012), 'The catechol-O-methyltransferase (COMT) Val158Met polymorphism moderates the effect of antenatal stress on childhood behavioural problems: longitudinal evidence across multiple ages', *Developmental Medicine & Child Neurology,* 54 (2), 148-54.
Thorndike, E. (1898), 'Some Experiments on Animal Intelligence', *Science,* 7 (181), 818-24.
Tidey, J. W. and Miczek, K. A. (1996), 'Social defeat stress selectively alters mesocorticolimbic dopamine release: an in vivo microdialysis study', *Brain Res,* 721 (1-2), 140-9.
Tunbridge, E. M. and Harrison, P. J. (2010), 'Importance of the COMT gene for sex differences in brain function and predisposition to psychiatric disorders', *Biological Basis of Sex Differences in ….*
Tunbridge, E. M., Weinberger, D. R., and Harrison, P. J. (2006a), 'A novel protein isoform of catechol O-methyltransferase (COMT): brain expression analysis in schizophrenia and bipolar disorder and effect of Val158Met genotype', *Mol Psychiatry,* 11 (2), 116-7.
Tunbridge, E. M., et al. (2004), 'Catechol-o-methyltransferase inhibition improves set-shifting performance and elevates stimulated dopamine release in the rat prefrontal cortex', *J Neurosci,* 24 (23), 5331-5.
Tunbridge, E. M., et al. (2012), 'The role of catechol-O-methyltransferase in reward processing and addiction', *CNS Neurol Disord Drug Targets,* 11 (3), 306-23.
Tunbridge, Elizabeth M, Harrison, Paul J, and Weinberger, Daniel R (2006b), 'Catechol-o-methyltransferase, cognition, and psychosis: Val 158 Met and beyond', *Biological psychiatry,* 60 (2), 141-51.
Vijayraghavan, S., et al. (2007), 'Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory', *Nat Neurosci,* 10 (3), 376-84.
Voon, V., et al. (2015a), 'Motivation and value influences in the relative balance of goal-directed and habitual behaviours in obsessive-compulsive disorder', *Transl Psychiatry,* 5, e670.

Voon, V., et al. (2015b), 'Disorders of compulsivity: a common bias towards learning habits', *Molecular Psychiatry,* 20 (3), 345-52.

Wager, Tor D. and Smith, Edward E. (2003), 'Neuroimaging studies of working memory', *Cognitive, Affective, & Behavioral Neuroscience,* 3 (4), 255-74.

Walt, Stéfan van der, Colbert, S. Chris, and Varoquaux, Gaël (2011), 'The NumPy Array: A Structure for Efficient Numerical Computation', *Computing in Science & Engineering,* 13 (2), 22-30.

Wanat, Matthew J., Bonci, Antonello, and Phillips, Paul E. (2013), 'CRF acts in the midbrain to attenuate accumbens dopamine release to rewards but not their predictors', *Nature neuroscience,* 16 (4), 383-85.

Watanabe, M., Kodama, T., and Hikosaka, K. (1997), 'Increase of extracellular dopamine in primate prefrontal cortex during a working memory task', *J Neurophysiol,* 78 (5), 2795-8.

White, Thomas P., et al. (2014), 'Sex differences in COMT polymorphism effects on prefrontal inhibitory control in adolescence', *Neuropsychopharmacology,* 39 (11), 2560-69.

Williams, G. V. and Goldman-Rakic, P. S. (1995), 'Modulation of memory fields by dopamine D1 receptors in prefrontal cortex', *Nature,* 376 (6541), 572-5.

Witte, Anja, et al. (2010), 'COMT Val158Met Polymorphism Modulates Cognitive Effects of Dietary Intervention', *Frontiers in aging neuroscience,* 2, 146.

Wood, Wendy and Rünger, Dennis (2015), 'Psychology of Habit', *Annual Review of Psychology,* 67 (1), 1-26.

Wunderlich, K., Smittenaar, P., and Dolan, R. J. (2012), 'Dopamine enhances model-based over model-free choice behavior', *Dopamine enhances model-based over model-free choice behavior*.

Xie, T., Ho, S. L., and Ramsden, D. (1999), 'Characterization and implications of estrogenic down-regulation of human catechol-O-methyltransferase gene transcription', *Mol Pharmacol,* 56 (1), 31-8.

Yavich, Leonid, et al. (2007), 'Site-Specific Role of Catechol-O-Methyltransferase in Dopamine Overflow within Prefrontal Cortex and Dorsal Striatum', *The Journal of Neuroscience,* 27 (38), 10196-209.

Yerkes, Robert M. and Dodson, John D. (1908), 'The relation of strength of stimulus to rapidity of habit-formation', *Journal of Comparative Neurology and Psychology,* 18 (5), 459-82.