Modelling animal conditioning with factored representations in dual-learning : explaining inter-individual differences at behavioural and neurophysiological levels

Florian Lesaint

A PhD Thesis by

Florian Lesaint

submitted to

Université Pierre et Marie Curie

École doctorale Informatique, Télécommunications et Électronique (Paris)

Institut des Systèmes Intelligents et de Robotique

in partial fulfillment of
the requirements for the degree of

Doctor in Computer Science

Modélisation du conditionnement animal par représentations factorisées dans un système d'apprentissage dual

Explication des différences inter-individuelles aux niveaux comportemental et neurophysiologique

Modelling animal conditioning with factored representations in dual-learning systems

Explaining inter-individual differences at behavioural and neurophysiological levels

26 September 2014

Committee

Angelo Arleo – ANC - UPMC/CNRS, FR	Examiner
Etienne Coutureau – INCIA - Univ. Bordeaux/CNRS, FR	Reviewer
Peter Dayan – Gatsby Computional Neuroscience Unit - UCL, UK	Examiner
Mehdi Khamassi – ISIR - UPMC/CNRS, FR	Supervisor
Arthur Leblois – CNPP - Univ. Descartes/CNRS, FR	Reviewer
Olivier Sigaud – ISIR - UPMC/CNRS, FR	Supervisor

Address

Institut des Systèmes Intelligents et de Robotique
Université Pierre et Marie Curie
ISIR - CNRS UMR 7222
Boite courrier 173
4 Place Jussieu
75252 Paris cedex 05 - France

Keywords

Computational neuroscience, Pavlovian conditioning, Instrumental conditioning, Reinforcement learning, Factored representations, Inter-individual variability.

Abstract

Pavlovian conditioning, the acquisition of responses associated to neutral stimuli that have been paired with rewards, and instrumental conditioning, the expression of a behaviour in order to achieve a goal, are at the heart of our learning capacities. However, while evidences clearly suggest that they are strongly entangled in the brain, they are mainly studied separately. The general framework of reinforcement learning (RL), learning by trials and errors to decide what to do in each situation to subsequently achieve a goal, while early used in the modelling of Pavlovian conditioning, is now mainly used for modelling instrumental conditioning. Models of Pavlovian conditioning rely now on more specific and dedicated architectures, focused on individual stimuli. This complicates the investigation of interactions between both types of conditioning since combining the various computational models is often neither straightforward nor natural. In the present thesis, we aim at finding key concepts that could be used in RL computational models to allow the study of Pavlovian conditioning, instrumental conditioning and their interactions. In particular, we model experimental data during autoshaping experiments in rats and negative automaintenance in pigeons.

When presented with a neutral lever before reward delivery, some rats come to approach, bite and chew the lever itself more and more avidly, whereas other rats come to approach the location of food delivery in a similar consumption-like manner. This inter-individual difference can be observed not only at the behavioural level but also at the physiological and pharmacological levels. When presented with a key light before reward delivery, pigeons start to peck more or less persistently at the key light. Furthermore, such pigeons persist in pecking the key light even if it blocks reward delivery. This maladaptive behaviour is more or less pronounced in pigeons. We show that combining a classical RL system, that learns values over situations, with a revised RL system, that learns values over individual features rather than classical canonic states and subsequently makes them compete to bias the behaviour towards reward-related stimuli, is sufficient to account for the aforementioned experimental data.

We explain maladaptive behaviours as the result of the detrimental collaboration of the two systems that, learning values over different elements, are not always guiding the behaviour towards an optimal solution. The model explains inter-individual differences as the result of a simple variation at the population level in the contribution and influence of each system on the overall behaviour. The model also explains some unexpected dopaminergic patterns with regard to the dominant hypothesis that dopamine parallels a reward prediction error signal, as the result of this signal being computed over features rather than over situations. Finally, we suggest that the revised version of the signal makes it also compatible with an alternative hypothesis that dopamine contributes to the acquisition of incentive salience, that makes reward-related stimuli wanted for themselves.

In conclusion, we present a unifying architecture able to explain yet unaccounted for experimental data at multiple levels, and show promising properties for the investigation of Pavlovian conditioning, instrumental conditioning and their interactions.

Acknowledgements

I wish to express my deep gratitude to the many people without whom this great scientific and human adventure would have never been possible, some of whom I shall thank many times for many different reasons.

First of all, I would like to thank Angelo Arleo, Arthur Leblois, Etienne Coutureau and Peter Dayan, who accepted to be part of my thesis committee, for taking the time to read my manuscript, for making useful comments and corrections, and for coming to Paris to attend my oral defence.

During these past three years of research, I had the chance to meet, discuss and collaborate with many great, intelligent and dedicated people. I would like to express my gratitude to all of them. I especially wish to thank Alain Marchand, Angelo Arleo, Emilio Cartoni, Etienne Coutureau, Gianluca Baldassarre, Mehdi Keramati, Peter Dayan, Quentin Huys and Tomasz Smolinski for their kind interest in my work, useful discussions and precious advice. I also wish to express my gratitude to Shelly Flagel, Jeremy Clark and Terry Robinson. Our collaboration was a real pleasure. I also wish to thank the many people I met during conferences, meetings or summer schools. Special thanks to Cati, Kun, Lena, Marta, Soledad and Felix with whom I shared endless nights of work, and great motivational and passionate discussions.

I am very grateful to my team colleagues for making ISIR such a nice place to work in, despite the summer heat, the permanent construction noise, the erratic functioning of the University, the overload of administration procedures, the poor quality food. . . Special thanks to Benoît, Emmanuel, J-B, Nicolas and Stéphane for their kindness, precious advice, useful help and discussions. I am of course grateful to my supervisors Olivier and Mehdi for guiding my steps through this dangerous land that is research. I must especially thank Olivier for his active presence, strong and regular interactions, and fixing so many typos in my writings. I am as well grateful to Mehdi for always pushing me forward, cheering me up, and being so optimistic and enthusiastic about my work.

I would especially like to thank my J01 office mates (and related), old and new, for the great times we had together. Many thanks to Ilaria for decorating the office with lots of useless boxes but also the most pleasant “*memorial*” sofa. Thanks also for making me discover all these crazy bio/zen/I ching stuff, always shining positive energy and entertaining us with spontaneous “*Che palle !*” amongst other Italian interjections. Many thanks to Serena for geeky discussions, funny linguistic situations, trying to teach us *important* Italian words, nice pasta parties, “C’est la drogue” food moments and culinary discoveries in general. Special thanks to Alain for entertaining J01 almost every single hour of every working day, especially given that his working desk was not in J01. Thanks also for nice brunches, laser games, the institution of the King cake marathon and deep learning preaches. Many thanks to Antoine for relaxing discussions, his active fight for improving our working conditions, and launching the Friday breakfast/ice-cream event. Many thanks to Charles for his perpetual cheerful mood, troll launches, Trololo disturbances, divisions by 37, sharing with us his projects and always setting a good ambiance. Many thanks to Charles, the other one, for making me discover medieval fencing and for often providing us with our daily dose of awkwardness. I should also include Ursine that to my opinion contributed as much as him to produce spontaneous strange faces and noises across the office. I am also grateful to Danesh for tolerating all these crazy annoying French talking guys, which includes me. Many thanks to Didier for providing our team with new questioning about really unexpected, diverse and/or passionate subjects. Many thanks to Erwan for nice board game breaks, interesting discussions about robotics and other casual subjects, making me promise to finish this stupid windowfarm project, always answering this irritating phone and repeated failed attempts to initiate some weekly after work event. Many thanks to Jean, the first, for his help on various matters, for providing me with some useful scripts, and perpetual cheerful interventions. Many thanks to Jean, the second, not so much for helping in the organization of club AMAC :), but for some great discussions about dopamine, kind of verbal bashing contests, big laugh sessions and some interesting guitar moments. Many thanks to Paul for geeky discussions, general advice about hardware, and his most helpful Sferes2 documentation. Many thanks to Sylvain for his kindness at my arrival in J01 and for teaching me some few tricks about verbal bashing (Yes, this is definitely compatible). Special thanks to Tony for his musical tutoring (sadly forgotten the day he left J01), for great parties and for trying to keep the whole team in shape and make our slogan be “Mens sana in corpore sano”. I must extend my gratitude to Audrey for making her flat almost an annexe of J01. I shall not forget the new comers, Arthur, Guillaume and Nassim, who kept Alain away from my peaceful desk, who kept playing CS games despite being such easy targets, who freed me from the organization of club AMAC, and for troll discussions. This, of course, is not an exhaustive list of all the stuff that made my days and kept me sane (or the opposite?) during this great adventure. I am also grateful to them for bearing my lousy jokes, my occasional spontaneous verbal bashing, my self-addressed mumbles and my permanent stressed attitude during the last few months writing this dissertation.

I also must thank all my former colleagues at LRDE and the Vaucanson team that set me on to the dark path of research. Although I have completely changed of research subject, my computer skills, curiosity, and rigour come from my time at LRDE. Special thanks to Alexandre for his early support and great recommendation letters.

My three year’s journey would not have been so enjoyable without my wonderful friends, who supported me, dealt with my busy schedule, had to bear my changing moods, accepted my mental absences, and provided me with an endless source of entertainment and joyful moments. Special thanks to Cécile and Cédric, who talked me out of giving up (too soon) our late drawing sessions at the Orsay museum, to whom I must add Marion, Arthur and Grégor for crazy board game nights. Many thanks to them, Fabienne, Muriel, Alban, Aymeric and Maxime for short but wonderful and refreshing winter holidays. I am also grateful to them and also Alexandra, Amélie, Anaïs, Anne-Pauline, Chloé, Claire, Clémence, Danielle, Elise, Elodie, Jeanne, Bertrand, Bruno, Colin, Guillaume, Laurent, Maxime, Sylvain, Thierry and Vivien for numerous, great and enjoyable moments of my daily life, and Mélanie for much more than just moments. I shall not forget my friends from abroad who kept in touch with me despite my sparse availability, maintained my poor English level, visited me or welcomed me in their places during short cheerful and relaxing periods: Barbara, Beatrix, Fulya, Karolina, Kun, Melis, Oksana, Soledad, Susi, Yasuyo, Yuka, Martijn and Paulo. Special thanks to Ana, Pedro and Lord Simon of Glencoe for such great time spent travelling across Europe. You are such wonderful friends.

And last but not least, many thanks to my parents for their love and constant support, even regarding my crazy choices in life, such as doing a PhD in computational neuroscience in a robotic laboratory after graduating from an engineering school in computer science and taking a one year's break in law school. You are awesome.

Contents

List of Figures

List of Tables

Résumé étendu

Introduction

Les neurosciences computationelles s'appuient sur des modèles informatiques pour étudier le système nerveux [Day94]. Elles permettent la synthèse entre les théories sur son fonctionnement et les larges quantités de données expérimentales récoltées à tous les niveaux, du niveau moléculaire au niveau comportemental. Nous nous intéressons ici aux capacités des animaux, humains compris, à apprendre, prédire, décider ou agir. Parmi ces capacités on distingue notamment le conditionnement Pavlovien et le conditionnement instrumental. Le conditionnement Pavlovien [Pav27] consiste en l'acquisition de réponses envers des stimuli neutres qui ont été associés avec des récompenses, comme par exemple quand nous salivons à la musique du marchand de glace. Le conditionnement instrumental [Ski38] consiste en l'acquisition d'un comportement dans le but d'atteindre un objectif, comme quand quelqu'un apprend à composer un numéro spécifique de téléphone pour appeler une personne. Combinés ensembles, ces conditionnements sont au cœur de nos capacités d'apprentissage et leur étude bénéficie de et repose grandement sur des modèles informatiques.

L'apprentissage par renforcement [SB98], qui modélise l'apprentissage par essai erreur pour décider quelle action choisir dans une situation donnée, est l'un des paradigmes les plus utilisés dans les modèles du conditionnent Pavlovien et instrumental. Comme exemple de son importante contribution dans le conditionnement Pavlovien, l'algorithme d'apprentissage *TD-Learning* [SB81], tout d'abord développé pour expliquer les capacités de prédiction des animaux dans certains tâches expérimentales [SB87; Sut88], a par la suite été prouvé reposer sur un signal qui pouvait être mis en parallèle avec l'activité des neurones dopaminergiques durant ces mêmes tâches [Sch+97]. Ce modèle permet donc de faire le lien entre l'expression d'un comportement et certains corrélats neuronaux sous-jacents. Il est maintenant communément accepté que le conditionnement résulte en partie de processus d'apprentissage par renforcement [Niv09; Mai09; DB12a; DN08].

Alors qu'ayant acquis sa notoriété dans l'étude du conditionnement Pavlovien [Sut88; Sch+97], le cadre moderne de l'apprentissage par renforcement est plus adapté à l'étude du conditionnement instrumental, où des actions sont en effet nécessaires pour atteindre un but. Les modèles informatiques récents du conditionnement instrumental sont souvent la résultante d'une combinaison de plusieurs systèmes d'apprentissage [Daw+05; Ker+11; Pez+13; DB12a; DD13]. Les modèles informatiques du conditionnement Pavlovien ne reposent principalement plus sur ce cadre général mais sur des architectures plus spécifiques et dédiées ou la notion d'action est secondaire [Sch+96; MM88; Cou+04; SM07; Jam+12; Has+10]. L'étude par des modèles informatiques des interactions entre ces deux types de conditionnement est limitée par la difficulté de combiner ces modèles, ceux-ci reposant sur des paradigmes et mécanismes rarement compatibles sans de profondes modifications.

Dans cette thèse, nous avons pour objectif de trouver des concepts clés utilisables dans le cadre de l'apprentissage par renforcement pour permettre l'étude des conditionnements Pavlovien et instrumental ainsi que de leurs interactions.

Observations

Apprentissage par renforcement

L'apprentissage par renforcement [SB98] est un cadre formel permettant l'étude et la résolution de problèmes de décision dans des environnements dont au moins la dynamique est inconnue a priori. Nous nous concentrons ici sur sa version standard, communément utilisée dans l'étude du conditionnement [Niv09; DB12a]. L'apprentissage par renforcement nécessite de définir la tâche à résoudre au travers d'un Processus de Décision Markovien (MDP), afin d'y faire tourner un algorithme qui cherche par essai erreur à en extraire une solution optimale.

Un MDP est défini par un ensemble $\langle S, Q, T, R \rangle$. S est un ensemble fini d'états qui représentent les différentes situations rencontrées par l'agent de manière abstraite (e.g. $s_0, \ldots, s_1$) perdant toute notion de similitude entre ces situations (e.g. la présence d'un levier). Nous nous intéressons cependant à une version factorisée de ces états dans laquelle ces informations sont préservées et accessibles aux algorithmes. A est l'ensemble fini d'actions possibles dans chaque état. $T : S \times A \times S \rightarrow [0, 1]$ est une fonction de probabilité de transition qui définit la probabilité $P(s'|s, a)$ d'atteindre l'état s' en faisant l'action a dans l'état s. $R : S \times A \rightarrow \mathbb{R}$ est une fonction de récompense qui définit la récompense $R(s, a)$ de faire l'action a dans l'état s. Cette formalisation en MDP implique

que l'état futur ne dépend que de l'état présent et de l'action réalisée, et non de l'historique de l'agent. Cela implique également que l'environnement reste stable (T et R invariants) au cour du temps. Ces contraintes sont souvent relâchées en neurosciences et l'intérêt des algorithmes d'apprentissage est de pouvoir s'y adapter au cours de l'expérience comme pourrait le faire un animal [Sut+92].

Les algorithmes d'apprentissage par renforcement n'ont pas besoin de connaissances a priori sur le problème pour pouvoir en trouver une solution optimale. Il en existe deux grandes catégories. Les algorithmes *Model-Based* (MB) construisent un modèle interne du monde à partir de leurs expériences, sur lequel il est possible d'inférer une solution optimale [Sut90; MA93; PW93; Glä+10; BT03; KS02; Wal+10; KS06; Bro+12]. Celle-ci est souvent représentée par une liste de valeur $\mathcal{Q}(s, a)$ pour chaque paire état-action qui définit l'espérance de gain cumulée si l'on réalise l'action a dans l'état s et que l'on suit ensuite le meilleur plan possible. Les algorithmes *Model-Free* (MF) ne passent pas par cette étape intermédiaire d'un modèle interne mais construisent directement ces valeurs en calculant et propageant un signal d'erreur de prédiction (RPE) entre ce qui est attendu et finalement observé, d'états proches de la récompense à ceux qui en sont éloignés [SB87; Sut88; Bar+83; WD92] . C'est ce signal qui a été corrélé à l'activité phasique de la dopamine [Sch+97; Gli11]. Même si les algorithmes MB et MF se comportent différemment au cours de l'apprentissage, ils convergent mathématiquement vers les mêmes solutions. Les systèmes MB sont souvent plus lents à exécuter, du fait d'un nombre potentiellement important de calculs par pas de temps, mais peuvent rapidement obtenir des politiques correctes, et s'adapter rapidement à tout changement dans la dynamique du monde. Les systèmes MF sont souvent rapides à l'exécution mais peuvent nécessiter un certain temps avant de trouver une solution acceptable ou réviser celle-ci en cas de changement de la dynamique du monde. Ces propriétés sont particulièrement importantes dans l'étude du conditionnement instrumental [AD81; Ada82].

Conditionnement

L'étude du conditionnement animal se divise principalement entre conditionnement Pavlovien [Pav27] et conditionnement instrumental [Ski38].

Le conditionnement instrumental résulte de la confrontation d'un animal à une tâche qui nécessite une séquence d'une ou plusieurs actions pour obtenir et maximiser des récompenses. Ce type de conditionnement permet d'étudier les stratégies mises en place par l'animal pour construire ces séquences et les réviser suite à une modification de leurs conséquences ou résultats [Dol+08; KH12]. Des expériences ont montré que l'animal ne se base pas toujours sur la même stratégie et qu'il est possible de distinguer les comportements dits *habituels* de ceux *orientés vers un but* (GD) [AD81; Ada82]. Les algorithmes MF et MB de l'apprentissage par renforcement présentent des propriétés très similaires à ces stratégies [Daw+05; Ker+11; DD13]. Les comportements GD sont souvent modélisés par des algorithmes MB. En effet, les comportements GD sont souvent visibles en début d'expérience, quand il faut s'adapter, trouver rapidement la récompense quitte à dépenser beaucoup d'énergie, et où les animaux semblent également capables en cas de modification de l'environnement (e.g. fermeture d'une porte) de trouver immédiatement la nouvelle solution comme s'ils possédaient une connaissance du monde sur laquelle planifier. Les comportements habituels sont souvent modélisés par des algorithmes MF. En effet, Une fois que la tâche a été répétée de nombreuses fois, il est souvent observé que les animaux sont plus rapides à effectuer celle-ci, n'hésitent plus face à des choix et en cas de changement d'environnement sont lents à s'adapter. Aussi, le comportement instrumental est souvent modélisé par une combinaison de deux systèmes d'apprentissage par renforcement classique, l'un MB et l'autre MF [Daw+05; Ker+11; Pez+13; Huy+12; DB13; Ger+14; DB02; Glä+10]. Le critère qui fait qu'un système domine plutôt que l'autre dans une situation donnée n'est pas encore clairement défini (e.g. [Ott+13; Foe+06; Fau+05; Ger+14; Ker+11; Daw+05]).

Les éléments biologiquement importants (e.g. la nourriture) produisent automatiquement certaines réponses chez les animaux (e.g. saliver), alors que d'autres éléments sont complètement neutres (e.g. une lumière). Pourtant, si ces éléments neutres sont suivis d'une récompense de manière répétée, les animaux commencent à développer à leur apparition une réponse souvent similaire à celle induite par la récompense. Le conditionnement Pavlovien [Pav27] étudie ces réponses, particulièrement leurs formes, leurs intensités et leurs propriétés. Cette capacité de certains éléments, appelés alors stimuli conditionnés, à induire ces réponses est à l'origine de phénomènes complexes. Certains phénomènes (e.g. *second order conditioning* [RR72; HR75], *sensory preconditioning* [Bro39; RR72]) mettent en évidence que dans une chaîne de stimuli neutres amenant à une récompense, les réponses peuvent être propagées des éléments les plus proches de la récompense à ceux les plus éloignés dans le temps, d'une manière ressemblant à ce que réalisent les systèmes d'apprentissage par renforcement MF. D'autres phénomènes (e.g. *blocking* [Kam67], *overexpectation* [LN98; KM96; Res99], *overshadowing* [Rey61; Mac76]) montrent clairement que les stimuli peuvent rentrer en compétition dans la génération des réponses conditionnées. Les algorithmes d'apprentissage par renforcement classique basés sur des MDP avec des états abstraits sont incapables de rendre compte de cet élément. Ces mêmes phénomènes, ainsi que d'autres (e.g. *ABA renewal* [BR94; Bou04]), suggèrent aussi que ce conditionnement nécessite parfois une connaissance plus générale de la tâche, comme dans les systèmes d'apprentissage par renforcement MB. Au final, les modèles récents du conditionnement Pavlovien se sont éloignés du cadre général de l'apprentissage par renforcement pour se tourner vers des architectures plus spécialisées capables de rendre compte d'un grand nombre de phénomènes non détaillés ici [Sch+96; SL06; LS08; MM88; Den+01; SM07; Cou+04; GN12; AS12].

Bien que généralement étudiés séparément, certains phénomènes sont clairement identifiés comme résultant d'une interactions entre conditionnement instrumental et Pavlovien (*negative automaintenance* [WW69; San+06; DW77; GR73; Kil03; Woo+74; Loc+76; O'C79; GH72], *Pavlovian-Intrumental-Transfer* [Hol+10; Huy+14; CB05; CB11] et *Conditioned Reinforcement Effect* [Wil94b; Ski38; RF09]). La majorité des protocoles utilisés dans le conditionnement instrumental font intervenir des stimuli propices au développement de réponses dans le cadre du conditionnement Pavlovien, et il n'est pas invraisemblable de penser que d'autres phénomènes restent à découvrir ou à réinterpréter comme ne résultant pas exclusivement d'un seul type de conditionnement. Alors que de nombreux modèles sont développés pour rendre compte du conditionnement Pavlovien et d'autres aussi nombreux sont développés pour rendre compte du conditionnement instrumental, il n'existe pour l'instant que peu de modélisations de phénomènes évidents d'interactions [Day+06; Huy+11]. En effet, les modèles récents de conditionnement reposent sur des paradigmes relativement différents qui rendent leur combinaison non triviale si ce n'est impossible sans de profondes modifications. Les modèles du conditionnement Pavlovien décrivent essentiellement les variations d'intensité d'une unique réponse conditionnée, et sa propagation à d'autres stimuli est souvent définie par une valeur sur des associations entre stimuli et plus rarement sur les stimuli eux-mêmes. La notion d'action y est presque toujours absente. Les modèles du conditionnement instrumental, reposant sur le cadre de l'apprentissage par renforcement, se concentrent principalement sur les séquences d'actions, la forme et l'intensité de ces actions n'étant souvent pas traitées. De plus, la notion de stimuli est cachée par l'utilisation d'états abstraits dés lors que plusieurs stimuli sont présents simultanément.

Hypothèse de travail

Puisant dans la littérature du conditionnement Pavlovien et instrumental, y compris de leurs interactions, nous avons cherché à identifier des concepts clés qui permettraient de réunir leurs études respectives dans un cadre commun, en prenant pour base l'apprentissage par renforcement classique, adapté au conditionnement instrumental, et l'étendant à des notions qui semblent nécessaires pour rendre compte du conditionnement Pavlovien.

Pour cela, nous nous sommes fortement inspirés de données expérimentales non encore reproduites à l'aide de modèles informatiques [Fla+11b; RF09; WW69; San+06]. Les premières données expérimentales mettent en lumière la présence d'une forte variabilité entre individus dans une population de rats sur l'expression d'une réponse conditionnée [Fla+11b]. Cette variabilité est présente non seulement au niveau comportemental, mais aussi aux niveaux physiologique (i.e. des enregistrements de l'activité dopaminergique) et pharmacologique (i.e. dans l'effet que pouvait avoir l'injection de certains drogues) [SR12; Fla+11b; RF09]. Les autres données proviennent de pigeons qui persistent à produire une réponse inappropriée dans une protocole particulier où il faut apprendre à se refréner d'agir [WW69; San+06]. Ces expériences suggèrent la présence de plusieurs systèmes d'apprentissage dont l'interaction conduirait à favoriser des comportements non optimaux. De plus, les comportements observés semblent clairement indiquer le développement d'une motivation particulière envers des éléments de la tâche d'où la présence d'au moins un système traitant les éléments séparément du reste.

Forts de ces intuitions, nous avons cherché à développer un modèle reposant sur la combinaison de plusieurs systèmes d'apprentissage par renforcement pouvant s'appuyer sur des représentations factorisées et ainsi accéder et traiter certains éléments indépendamment de l'état où ils se trouvent. Nous avons cherché à valider ce modèle sur ces données et plus généralement à en discuter la portée pour l'investigation des interactions entre Pavlovien et instrumental.

Résultats

Modèle informatique

Le modèle informatique développé au cours de cette thèse repose sur l'architecture présentée en Figure 1. Il est composé principalement de deux systèmes d'apprentissage par renforcement distincts qui collaborent à la sélection d'actions à chaque pas de temps au cours de l'expérience. Un système favorise des comportement rationnels et optimaux pour maximiser les gains, l'autre amène à des choix plus impulsifs.

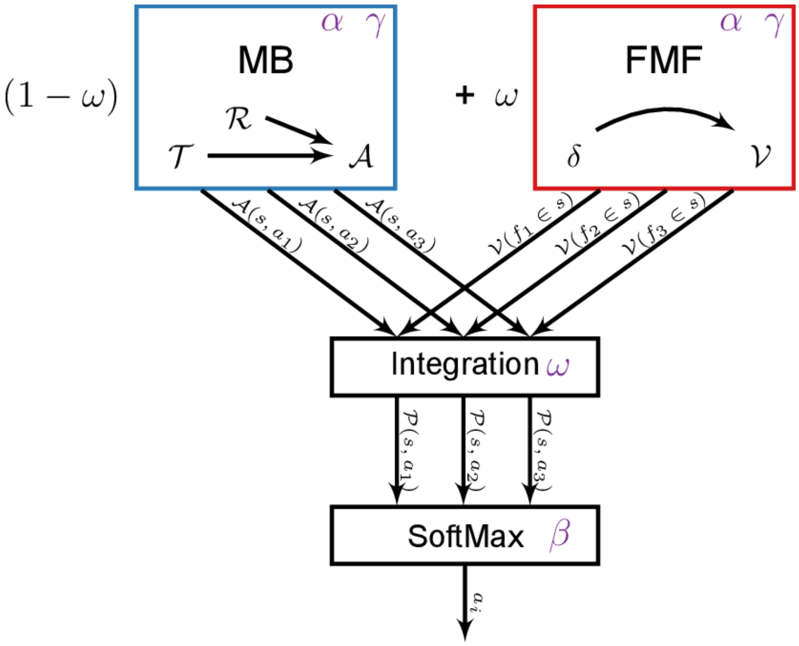


Figure 1 – Architecture générale du modèle. *Le modèle est composé d'un système* Model-Based *(MB, en bleu) et d'un système* Feature-Model-Free *(FMF, en rouge) qui fournissent respectivement une fonction Avantage $\mathcal{A}$ pour les actions a_i dans un état donné s et une fonction valeur $\mathcal{V}$ pour chaque élément f_i composant l'état courant. Ces valeurs sont intégrées en $\mathcal{P}$, avant d'être passées au mécanisme de sélection de l'action (softmax). Certains composants reposent sur des paramètres (en violet).*

Le premier système est un système *Model-Based* qui apprend les conséquences à long terme des actions en estimant un modèle approximatif du monde (une fonction de transition $\mathcal{T}$ et une fonction de récompense $\mathcal{R}$) à partir duquel il est possible d'anticiper les actions à réaliser, i.e. planifier. Par exemple, le modèle est suffisant pour anticiper la distribution d'une récompense et qu'il est donc intéressant de s'approcher de la mangeoire avant même de voir celle-ci tomber dedans. Dans notre implémentation, ce système construit la fonction Avantage $\mathcal{A}$ qui évalue à partir du modèle l'avantage de réaliser chaque action dans chaque situation, et qui est donnée par les formules

$$\mathcal{Q}(s,a) \leftarrow \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \max_{a'} \mathcal{Q}(s',a') \tag{1}$$

$$\mathcal{A}(s,a) \leftarrow \mathcal{Q}(s,a) - \max_{a'} \mathcal{Q}(s,a') \tag{2}$$

où le taux d'actualisation $0 \leq \gamma \leq 1$ représente la préférence pour obtenir la récompense immédiatement plutôt que retardée et $\mathcal{Q}(s,a)$ est la valeur estimée des gains futurs à faire l'action a dans l'état s (cela correspond à la récompense accumulée attendue en suivant le meilleur plan d'actions). A chaque pas de temps, l'action avec la valeur la plus élevée est celle qui permettra d'accumuler le plus de récompenses le plus tôt possible sur le long terme (e.g. s'approcher d'une mangeoire pour manger la nourriture dés qu'elle y tombe). L'équation 1 représente le processus par lequel un agent estime les conséquences futures de réaliser l'action a dans l'état s. Si l'action a est supposée amener à une récompense $\mathcal{R}(s,a)$ ou avec une bonne probabilité $\mathcal{T}(s'|s,a)$ à un autre état s' avec une haute probabilité d'action $\mathcal{Q}(s',a')$ alors l'agent lui associe une forte $\mathcal{Q}$-valeur. L'équation 2 déduit l'avantage de réaliser une action a dans un état s en comparant les Q-valeurs à celles de toutes les autres actions possibles dans l'état. Il est à noter que d'autres implémentations pourraient être envisagées [Glä+10; BT03; KS02; Wal+10; KS06; Bro+12; Sut90; MA93; PW93].

Le second système est un système *Model-Free*. Il n'apprend pas de modèle interne du monde mais apprend progressivement à associer une valeur à chaque élément de l'environnement, favorisant les actions vers ceux les plus valorisés. En conséquence, ce système produit un comportement réactif, similaire aux habitudes [Gra08; DD13].

Dans l'apprentissage par renforcement traditionnel (e.g. le système MB), les valeurs sont généralement apprises sur les états et non les éléments qui les composent, de telle sorte que les similarités entre états (e.g. la présence d'une mangeoire) sont ignorées. Le système actuel apprend des valeurs $\mathcal{V}$ sur les éléments (e.g. un levier, de la nourriture) et est appelé *Feature-Model-Free* (FMF). L'apprentissage progressif des valeurs repose sur une erreur de prédiction de récompense (RPE) δ, utilisée comme suit :

$$\begin{gathered} \mathcal{V}(f) \leftarrow \mathcal{V}(f) + \alpha\delta \\ \delta \leftarrow r + \gamma \max_{f' \in s'} \mathcal{V}(f') - \mathcal{V}(f) \end{gathered} \tag{3}$$

où f est l'élément sur lequel se concentre l'action a dans l'état s. Le max suggère que tous les éléments f' qui composent l'état suivant s' sont considérés et que l'élément le plus valorisé est utilisé pour calculer la RPE, même s'il ne sera pas forcément l'élément sur lequel se focalisera la prochaine action [WD92]. On fait l'hypothèse traditionnelle que ce signal d'erreur correspond à l'activité phasique dopaminergique [Sch98; Gli11]. Ce signal permet de mettre à jour et d'attribuer des valeurs, vues comme source de motivation, à des éléments sans la nécessité d'un modèle interne du monde. Quand un élément est complètement attendu, il ne devrait pas y avoir de RPE car sa valeur est complètement anticipée ; si un évènement est surprenant, la RPE sera positive. Ces valeurs apprises biaisent le comportement vers les actions qui se focalisent sur les éléments les plus valorisés. Cela peut conduire à des comportements sous-optimaux en ce qui concerne l'accumulation de récompenses. Le système FMF modélise l'attraction développée par certains stimuli associés aux récompenses, i.e. le phénomène d'*incentive salience* [MB09; DB12b; Ber07].

Le modèle ne base pas sa décision sur un seul système à la fois. Les valeurs du système MB ($\mathcal{A}_{MB}$) et du système FMF ($\mathcal{V}_{FMF}$) sont intégrées de telle sorte qu'une seule décision est prise à chaque pas de temps. Les valeurs sont combinées par une somme pondérée et transmises à la fonction *softmax*, un mécanisme de sélection d'action qui les convertit en probabilités de choisir ces actions dans une situation donnée (Figure 1). L'intégration est réalisée comme suit :

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{MB}(s,a) + \omega \begin{cases} 0 & \text{si } a = ngo \\ \mathcal{V}_{FMF}(f) & \text{avec } f = c(s,a) \text{ sinon} \end{cases} \tag{4}$$

où $0 \leq \omega \leq 1$ est un paramètre de combinaison qui définit l'importance de chaque système dans le comportement généré par le modèle. La fonction d'éléments $c : \mathcal{S} \times \mathcal{A} \rightarrow \{\text{touche(s)}, \text{magasin}, \text{nourriture}, \emptyset\}$ retourne l'élément sur lequel se concentre l'action a dans l'état s (e.g. elle retourne la touche lumineuse quand l'action est de s'engager vers cette touche). Nous faisons l'hypothèse que seules les actions d'engagement vers un élément (e.g. approche ou interaction) bénéficient du bonus calculé par le système FMF, d'où la condition sur l'action $a = ngo$ dans la seconde partie de l'équation. Cette hypothèse se base sur les travaux de [GM+12; GM+14] qui suggèrent la présence d'un biais pour s'engager par rapport à se refréner.

Le modèle apprend par expérience à chaque pas de temps. Les systèmes FMF et MB sont mis à jour par rapport à l'action a choisie par le modèle complet dans l'état s, et l'observation de l'état s' et la récompense r en résultant.

Nous utilisons des algorithmes évolutionnaires [Deb+02; MD10] pour optimiser les paramètres du modèle à chacune des études de conditionnement que nous souhaitons simuler, i.e. minimiser la différence entre les résultats expérimentaux et ceux simulés au niveau du comportement, et ainsi obtenir des résultats qualitativement proche des originaux.

Variabilité inter-individuelle dans une tâche d'autoshaping

Utilisant une procédure de conditionnement Pavlovien, où un levier est présenté pendant 8 secondes, suivi immédiatement après par une distribution de nourriture dans un magasin adjacent, Flagel et al. [Fla+11b] ont observé une forte variabilité dans les réponses induites chez des rats. Certains rats (*sign-trackers*; ST) développent rapidement une réponse d'approche suivi du grignotement du levier, bien que cela ne soit pas nécessaire à la distribution de nourriture. D'autres rats (*goal-trackers*; GT), développent une approche similaire mais vers la mangeoire (Figure 2). Ils ont de plus constaté que l'activité dopaminergique enregistrée dans le cœur du nucleus accumbens différait selon les rats. Chez les ST, cette activité est conforme à la littérature actuelle sur le sujet, avec un pic d'activité à l'apparition de la nourriture qui au cours de temps se déplace à l'apparition du levier [Sch+97]. De plus, toute interruption du fonctionnement de la dopamine (e.g. par l'injection de drogues) empêche le développement de ce type de comportement. Chez les GT, au contraire, l'activité est différente, le pic de dopamine n'est pas propagé et un second apparaît à l'apparition du levier. De plus, l'acquisition du comportement ne semble pas sensible au blocage de la dopamine.

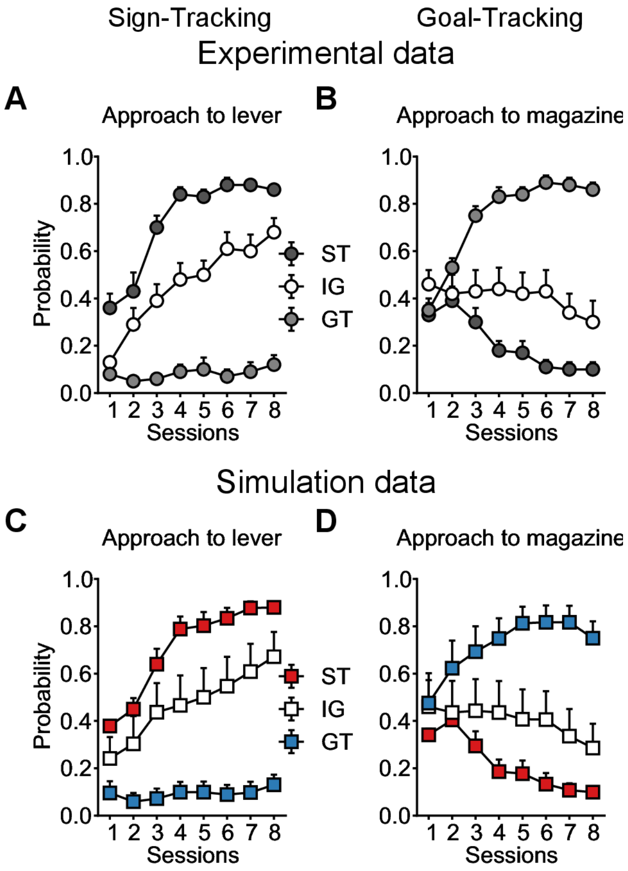
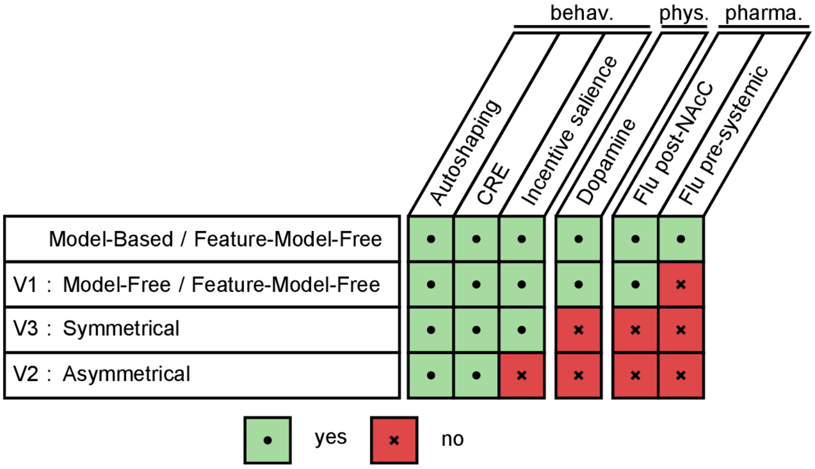


Figure 2 – Reproduction des différences de comportement dans une population de rats réalisant une expérience d'*autoshaping*. *Probabilité moyenne de s'engager au moins une fois avec le levier* **(A,C)** *ou le magasin* **(B,D)** *pendant les essais. Les données sont exprimées en moyennes ± SEM et regroupées par blocs de 50 essais (2 sessions).* **(A,B)** *Reproduction des résultats expérimentaux de Flagel et al. [Fla+09] (Figure 2 A,B). Les* sign-trackers *(ST) appuient le plus sur le levier (noir), les* goal-trackers *(GT) appuient le moins sur le levier (gris), et le groupe intermédiaire (IG) est entre les deux (blanc).* **(C,D)** *Simulation de la même procédure (carrés) avec le modèle informatique. Les rats simulés sont répartis entre ST ($\omega = 0.499$) en rouge, GT ($\omega = 0.048$) en bleu et IG ($\omega = 0.276$) en blanc. Le modèle reproduit les même différences comportementales. Avec de l'entraînement, les ST s'engagent de plus en plus vers le levier et de moins en moins vers le magasin, alors que les GT négligent le levier au profit d'un engagement croissant vers le magasin. Les IG oscillent entre les deux.*

Ces données posent problème aux modèles actuels du conditionnement Pavlovien [Cla+12] qui ne rendent souvent compte que d'un seul type de comportement à la fois et d'un seul type de réponse. Elles posent également problème à d'autres modèles qui s'appuieraient essentiellement sur le cadre de l'apprentissage par renforcement classique. Dans [Les+14b] nous comparons notre modèle à ces modèles plus classiques et montrons qu'il est le seul à même de rendre compte des diverses données expérimentales collectées autour de cette expérience [Fla+09; Fla+11b; Fla+07; RF09; Mey+12] (Figure 3).



	behav.			phys.	pharma.	
	Autoshaping	CRE	Incentive salience	Dopamine	Flu post-NAcC	Flu pre-systemic
Model-Based / Feature-Model-Free	●	●	●	●	●	●
V1 : Model-Free / Feature-Model-Free	●	●	●	●	●	×
V3 : Symmetrical	●	●	●	×	×	×
V2 : Asymmetrical	●	●	×	×	×	×

● yes × no

Figure 3 – Synthèse des simulations. *Chaque ligne représente un modèle différent composé d'une paire de systèmes d'apprentissage par renforcement. Chaque colonne représente une expérience simulée. Les expériences sont groupées par le type de données expérimentales qu'elles impliquent : comportementales (*autoshaping *[Fla+09; Fla+11b], CRE [RF09],* Incentive salience *[MB09; DB12b]), physiologique [Fla+09] et pharmacologique (Flu post-NAcC [SR12],* Flu pre-systemic *[Fla+09]).*

Nous modélisons l'expérience décrite avec le MDP illustré en Figure 4. L'agent commence dans un état vide (s_0) où il n'y a rien d'autre à faire qu'explorer. A un certain moment, le levier apparaît (s_1) et l'agent doit faire un choix : il peut s'engager avec le levier (s_2) et interagir avec, s'engager vers le magasin (s_4) et interagir avec, ou continuer à explorer (s_3, s_6). A un certain moment, le levier est rétracté et de la nourriture est délivrée. Si l'agent est loin du magasin (s_5,s_6), il doit d'abord s'en approcher. Une fois prêt (s_7), il mange la nourriture. Il finit dans un état vide (s_0) qui symbolise le début de la période de repos entre deux essais (ITI) : pas de nourriture, pas de levier et un magasin toujours présent mais vide. Le même MDP et le même jeu de paramètres est utilisé pour toutes les expériences, i.e. notre modèle peut faire le lien entre des données physiologiques et l'expression d'un comportement.

Aussi, notre modèle explique la différence entre les ST et le GT par une pondération différente dans la contribution des systèmes MB et FMF (Figure 2). Les ST se reposent principalement sur le système FMF, tandis que les GT se reposent principalement sur le système MB. De plus, cette différence se retrouve au niveau du signal de prédiction d'erreur δ du système FMF que l'on peut alors de nouveau mettre en parallèle avec les données dopaminergiques, et ainsi résoudre le conflit que ces données présentaient par rapport à la littérature actuelle. Nous faisons la proposition que certains signaux dopaminergiques peuvent encoder une erreur de prédiction calculée sur des stimuli individuels

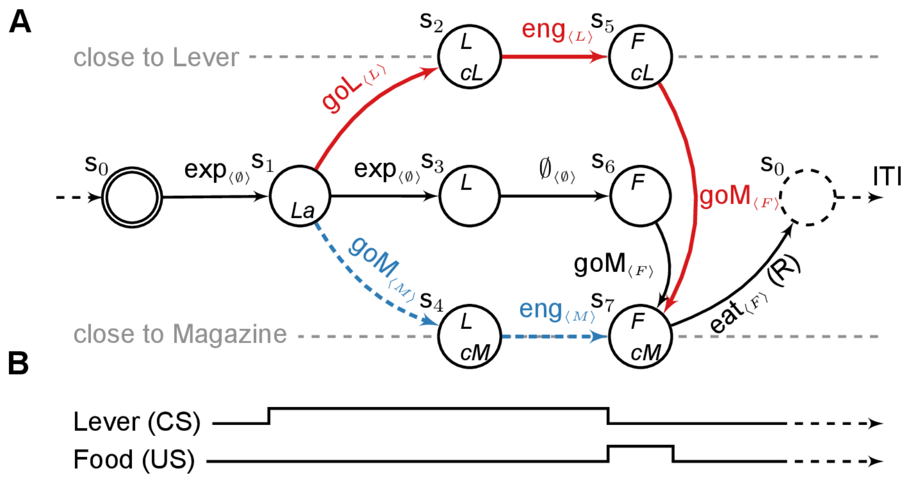


Figure 4 – Représentation informatique de la procédure d'*autoshaping* (A) *MDP représentant l'expérience décrite en [Fla+09; Fla+11b; RF09; Mey+12]. Les état sont décrits par un jeu de variables : L/F - levier/nourriture disponible, cM/cL - proche du magasin/levier,La - apparition du levier. L'état initial est doublement cerclé, l'état final est représenté en pointillé et termine l'épisode courant. Les actions sont : s'engager avec le stimulus le plus proche (eng), explorer (exp), s'approcher (go) du magasin ou du levier ou manger (eat). Pour chaque action, l'élément central à l'action est présenté entre crochets. Le chemin que les ST devraient préférer est en rouge. Le chemin que les GT devraient préférer est en bleu pointillé.* **(B)** *Chronologie correspondant au déroulement du MDP.*

et non des situations globales. Pour finir, cette différence de contribution dans les deux systèmes explique également pourquoi le comportement des GT ne semble pas affecté par l'injection de flupentixol, un antagoniste de la dopamine, alors que celui des ST est bloqué tant dans son acquisition que son expression.

Comportements inadaptés dans une tâche de *negative automaintenance*

Dans une procédure presque similaire à celle utilisée par [Fla+11b] mais avec une touche lumineuse à la place du levier, et tout contact avec cette touche entraînant une omission de la distribution de la nourriture, Williams and Williams [WW69] ont montré que certains pigeons étaient incapables de s'empêcher de becqueter cette touche malgré son effet néfaste. De plus, d'autres résultats semblent indiquer que tous les pigeons ne sont pas incapables d'apprendre à retenir leurs coups de becs [San+06], suggérant ici aussi une variabilité dans les réponses observées pour une même tâche dans différent individus de la même espèce. Dans l'expérience de Williams and Williams [WW69], divers protocoles mettent en valeur que, bien qu'incapables de se refréner de becqueter, les pigeons sont capables dans une certaine mesure de choisir vers quel élément saillant ils dirigent leurs coups de bec.

Ces données posent une fois de plus problème à la littérature classique, avec plusieurs réponses tournées vers plusieurs cibles, et une variabilité inter-individuelle difficilement explicable par un seul système. Dans [Les+14a] nous confirmons que notre modèle est capable d'expliquer également ce jeu de données sans modifications.

Le problème est presque identique à celui sur la tâche d'*autoshaping* de Flagel et al. [Fla+11b] (Figure 5). Nous avons remplacé dans le MDP précédent le levier par une touche lumineuse, et rajouté la possibilité, ici nécessaire, de se réfréner d'agir (ngo) une fois proche d'un stimulus. Le contact (eng) avec la touche lumineuse amenant directement au début de la période de repos (ITI) sans obtention de la récompense. D'autres MDP similaires ont été utilisés pour les différents protocoles, en rajoutant d'autres chemins pour d'autres leviers.

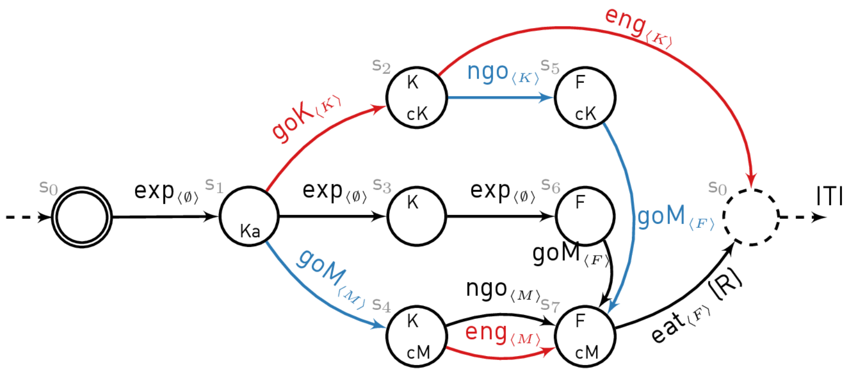

eng$_{\langle K \rangle}$
s_2
ngo$_{\langle K \rangle}$
s_5
K
cK
F
cK
goK$_{\langle K \rangle}$
s_0
exp$_{\langle \emptyset \rangle}$
s_1
Ka
exp$_{\langle \emptyset \rangle}$
s_3
K
exp$_{\langle \emptyset \rangle}$
s_6
F
goM$_{\langle F \rangle}$
s_0
ITI
goM$_{\langle F \rangle}$
goM$_{\langle M \rangle}$
s_4
ngo$_{\langle M \rangle}$
s_7
K
cM
eng$_{\langle M \rangle}$
F
cM
eat$_{\langle F \rangle}$
(R)

Figure 5 – Représentation informatique de la procédure de negative automaintenance. *MDP représentant l'expérience 1 de Williams and Williams [WW69] et du protocole* Brief PA *de Sanabria et al. [San+06]. Les états sont représentés par un jeu de variables : K/F - touche lumineuse négative/nourriture est disponible (le magasin est toujours disponible, même s'il n'est pas montré), cM/cK - proche du magasin/touche lumineuse, Ka - apparition de la touche lumineuse. L'état initial est doublement cerclé, l'état terminal est représenté en pointillé et termine l'épisode courant. Les actions sont : s'engager (eng) ou se réfréner de s'engager (ngo) avec le stimulus le plus proche, explorer (exp) ou s'approcher (go) du magasin ou de la touche de lumière, et manger (eat). Pour chaque action, l'élément central à l'action est présenté entre crochets.*

Notre modèle arrive à reproduire les données tant de Williams and Williams [WW69] que Sanabria et al. [San+06] avec le même jeu de paramètres, à l'exception principale de la pondération entre les deux systèmes (Figure 6). Ainsi, nous expliquons l'incapacité des pigeons à s'empêcher de becqueter par une forte prépondérance du système MF dans la décision. A l'inverse, les pigeons capables de ne pas toucher la lumière ont une prépondérance pour le système MB. Notre modèle explique également la prévalence des stimuli contingents aux récompenses à attirer les coups de becs par rapport à ceux présents tout au long de l'expérience, même dans le cas où l'interaction qui en résulte est néfaste à l'accumulation de récompenses.

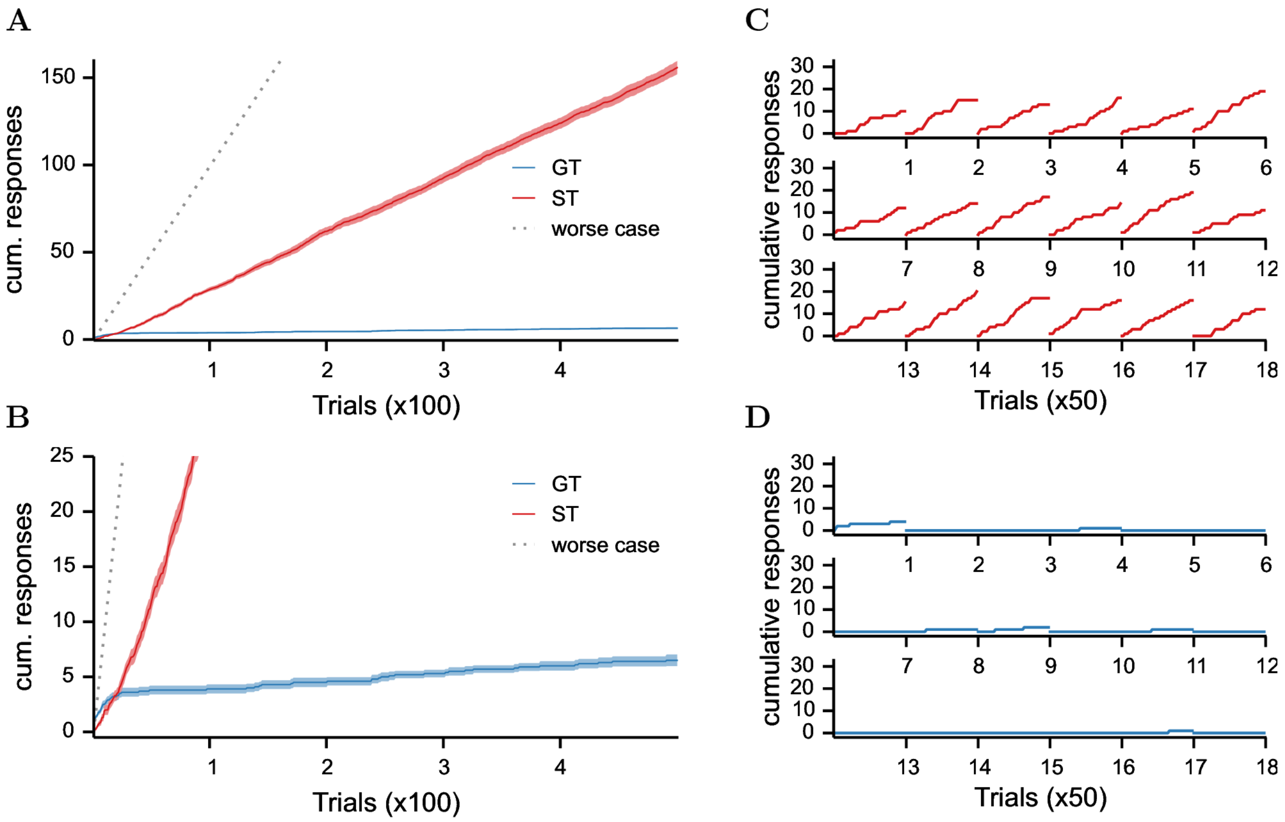


Figure 6 – Simulation de l'expérience 1 de Williams and Williams [WW69] et du protocole *Brief PA* de Sanabria et al. [San+06]. (A) *Coup de becs cumulés envers la touche lumineuse négative réalisés par 8 pigeons GT simulés (bleu) et 8 pigeons ST simulés (rouge). La courbe en gris pointillé simule le pire scénario (si les pigeons avaient becqueté la touche lumineuse à chaque essai). Les données sont exprimées en moyenne ± SEM.* **(B)** *Agrandissement de (A) pour un meilleure lisibilité de la courbe bleue (GT).* **(C)** *Coups de becs cumulés pour un pigeon ST par bloc de 50 essais. A mettre en parallèle avec la Figure 1 de Williams and Williams [WW69].* **(D)** *Coups de becs cumulés pour un pigeon GT par bloc de 50 essais.*

De manière plus générale, nous expliquons une fois de plus la variabilité des comportements par la seule différence de pondération entre deux systèmes MB et FMF. Nous mettons en évidence que cette collaboration peut effectivement amener à des comportements inadaptés.

Discussion

Synthèse des contributions

Cette thèse présente nos contributions dans la modélisation de certains phénomènes Pavloviens et instrumentaux à l'aide du cadre de l'apprentissage par renforcement, étendu pour utiliser des représentations factorisées.

Nous expliquons que les variabilités inter-individuelles observées chez les rats [Fla+11b] et les pigeons [WW69; San+06] sont la résultante d'une collaboration de deux systèmes basés sur des principes d'apprentissage par renforcement [Les+14b; Les+de]. Le premier système (MB) évalue les situations rencontrées de manière globale alors que le second système (FMF) traite des éléments indépendants. Ces systèmes ne favorisent donc pas toujours les mêmes actions, et selon la prépondérance accordée à l'un plutôt qu'à l'autre, différents individus peuvent présenter des comportements radicalement différents sur le long terme [Fla+07; Fla+09; RF09; Fla+11b; MB09; DB12b; SR12; Mey+12; Mey+14].

Cette collaboration peut amener à des comportements inadaptés [BB61; Her86; GM+12], comme ceux étudiés chez les pigeons [Les+de]. Nous montrons que l'acquisition d'une certaine valeur, source de motivation, par des éléments contingents aux récompenses, peut biaiser de manière permanente le comportement vers des actions néfastes à l'accumulation de récompenses.

Le calcul de la RPE au niveau du système FMF dépend du comportement au niveau du modèle, et peut donc être différente d'un individu à un autre. Nous expliquons ainsi que l'incohérence observée entre l'activité dopaminergique des GT [Fla+11b] et ce qui est attendu par la littérature classique [Sch+97] viendrait de la différence de comportement induite par la différence de pondération entre les deux systèmes et du calcul de la RPE sur des éléments individuels et non sur la situation globale à laquelle les rats sont confrontés.

Pour finir, notre modèle explique l'acquisition d'*incentive salience* [Ber07; MB09; Ber12] par l'évaluation individuelle de certains éléments saillants. L'attribution d'une valeur propre pour ces éléments, permet alors de biaiser le comportement en favorisant des actions d'approches et d'engagement avec ces derniers.

Perspectives et limites

Jusqu'à présent, nous avons uniquement envisagé l'utilisation de représentations factorisées dans le système MF de notre architecture. Cependant, rien n'empêcherait d'utiliser de telles représentations dans le système MB. Il serait même surprenant qu'un système se passe de représentations plus riches si elles sont accessibles. Leur utilisation pourrait cependant être différente dans chaque système. Il existe déjà un certain nombre d'algorithmes MB s'appuyant sur des représentations factorisées [Bou+00; Deg+06; VB08]. Ceux-ci utilisent leur connaissance de la structure du monde pour représenter les fonctions valeurs de manière plus compacte, généraliser certains calculs et ainsi gagner en espace et temps de calculs requis. Ces optimisations ne servent qu'à étendre les algorithmes classiques à de plus gros problèmes, mais ne changent en rien les solutions optimales trouvées. Aussi, il serait possible de remplacer notre système MB par une version factorisée traditionnelle sans changer les résultats. Cependant, les propriétés de généralisation de ces algorithmes peuvent introduire des comportements en début d'apprentissage qui diffèrent des algorithmes classiques, et il serait opportun d'étudier si de tels comportements peuvent être observés chez certains animaux et confirmer l'utilisation de tel algorithmes. Il est aussi à noter qu'à notre connaissance, il n'existe pas de version factorisée des systèmes MF classiques, qui n'ont pas accès à la structure du monde pour en extraire une organisation compacte des valeurs.

Bien que notre modèle soit la combinaison d'un système MB et d'un système MF, il est important de le distinguer des modèles du conditionnement instrumental qui impliquent aussi ces deux aspects [Daw+05; Ker+11; Pez+13], où des études qui suggèrent également la présence de ces deux composantes dans le Pavlovien [Jon+12; RB13]. Dans le cas instrumental, la combinaison de deux modèles est utilisée pour reproduire la variation des capacités d'un même individu selon certains critères ou expériences (e.g. motivation ou entraînement) [Ott+13; Daw+05; Ker+11; Daw+11; Dol+12; Bei+11]. Dans le cas Pavlovien, il est également question de rendre compte de la capacité d'un même individu à prendre en compte certaines informations passées dans une nouvelle situation [RB13; Jon+12]. Notre modèle au contraire, utilise cette combinaison pour rendre compte de la variabilité de comportement de différent individus dans les mêmes conditions théoriques. Nous pensons avoir ici principalement modélisé les aspects MF du Pavlovien et MB de l'instrumental. Il est à noter que le système FMF se comporte comme un système MF classique dans certains cas (par exemple, pour les ST, il se comporte comme attendu par la théorie classique [Sch+97; Sch98; Sut88]) et on pourrait envisager de l'y substituer dans les modèles utilisant des systèmes MF classiques. Un modèle plus complet pourrait intégrer les quatre systèmes (MF/MB instrumental et Pavlovien), en s'inspirant des modèles déjà existants, soulevant la question alors importante de savoir comment s'organise leur intégration [Yin+08; Mee+12; LO12; Mee+10; Mai09].

Tous les algorithmes MF reposent sur un signal d'erreur de prédiction, mais ce signal n'est pas forcément calculé de la même manière [SB87; Sut88; WD92; SB98]. Bien que l'hypothèse que la dopamine encode effectivement un signal d'erreur reste assez majoritaire dans la communauté [Sch+97; Sch10; Sch13; Mor+06; Roe+07; Eno+11], il existe un débat quant à son mode de calcul. Certaines études suggèrent un calcul de type "Q-Learning" i.e. on considère que la prochaine action réalisée sera celle avec la valeur la plus élevée [Roe+07], d'autre un calcul de type "SARSA", i.e. au moment du calcul de la RPE, l'action suivante est déjà choisie et donc prise en compte [Mor+06], cette question est en cours d'investigation et il semble que le problème est plus complexe [Bel+12a; Bel+13]. Nous avons arbitrairement choisi une règle de type "Q-Learning" mais l'autre possibilité reste à explorer. De manière plus générale, la RPE calculée dans notre système FMF repose sur la propagation de valeur entre éléments et non entre états, il n'est donc pas impossible que les données conflictuelles de Morris et al. [Mor+06] et Roesch et al. [Roe+07] puissent résulter de cette différence plus que le type de règle impliqué, car les protocoles de ces deux expériences ne sont pas identiques au niveau de la présentation des stimuli conditionnés. Pour finir, les données de Roesch et al. [Roe+07] suggèrent que l'acquisition du comportement est plus rapide que l'évolution de la dopamine. Notre modèle ne dépendant que partiellement sur le système FMF, en complètement d'un autre système MB supposé plus rapide, pourrait expliquer ces données. Nous n'avons pas encore étudié la dynamique de notre modèle ni les données exploitées sur cet aspect.

Une partie importante de nos résultats s'appuie sur l'hypothèse que la présence d'un stimulus pendant la période de repos entre les essais induit une diminution de la valeur qu'il a pu acquérir pendant la phase de conditionnement [Les+14b; Les+14a; Les+de]. Cette hypothèse explique pourquoi la mangeoire ne serait pas aussi attractive qu'un levier [Fla+11b; RF09], ou qu'une touche lumineuse qui ne s'éteint jamais n'arriverait pas à attirer les coups de becs des pigeons en présence d'une autre touche lumineuse contingente à la distribution de nourriture [WW69]. Des travaux préliminaires semblent montrer que cette dévaluation n'est pas dûe à un quelconque engagement envers ces éléments pendant cette période. Nous avons proposé dans [Les+14a] des protocoles qui permettraient de confirmer que la simple présence en serait la cause. Il serait également possible d'envisager que chaque stimulus diffère dans la valeur qu'il est capable d'acquérir, en fonction de sa forme ou d'autres propriétés [RW72; Mey+14; Hol77], comme certains phénomènes Pavlovien semblent le suggérer [LN98; Rey61; Mac76; Kam67; KM96; Res99].

Pour finir, nous avons pris le parti d'utiliser l'apprentissage par renforcement, utilisé dans le conditionnement instrumental, comme base de nos travaux et de l'étendre au concept des représentations factorisées présent et nécessaire au cadre Pavlovien. Une autre approche aurait pu être de prendre pour base un modèle du conditionnement Pavlovien, comme par exemple les travaux sur la *Latent Cause Theory* [Cou+04; Cou+06; GN12] et de l'étendre avec des notions nécessaires au conditionnement instrumental comme les actions. Cette approche a notamment été utilisée par Cartoni et al. [Car+13] pour rendre compte de certains phénomènes d'interactions entre instrumental et Pavlovien. Dans les deux cas, cette approche soulève la question de la différence entre réponses/réflexes conditionnées et actions volontaires [Bal+08].

Conclusion

Cette thèse est une petite étape dans le développement d'un cadre unifié pour l'étude des conditionnements Pavlovien et instrumental, et particulièrement pour l'étude des tâches expérimentales qui les impliquent tous les deux. En prenant inspiration de différences inter-individuelles et de comportements inadaptés observés dans le conditionnement des rats et des pigeons, nous avons pu extraire deux concepts qui semblent importants pour ce conditionnement : la combinaison de plusieurs systèmes d'apprentissage par renforcement et le traitement individuel des stimuli ainsi que leur compétition. L'interprétation d'autres données expérimentales pourrait bénéficier de cette approche. Nous espérons que poursuivre cette démarche apportera de nouvelles idées dans le domaine.

Chapter 1

Introduction

The best material model of a cat is another, or preferably the same, cat. - Norbert Wiener

1.1 Motivations

Computational models are determinant tools in the scientific study of the nervous system [Day94]. They help to synthesize large quantities of empirical data in all disciplines of neuroscience, from studies at the molecular level to studies at the behavioural level. Theories regarding the mechanisms and functional roles of the various elements of the nervous system, e.g. anatomical parts or specific chemicals, or even more general capacities, e.g. memory, are often validated afterwards, suggested beforehand and/or formalized by computational models. They allow one to replicate results, to explain findings with simple notions, to draw predictions and to guide research processes towards important questions.

Computational neuroscience is of particular interest for studying how one learns from its interactions with the world, anticipates future events and ultimately selects and/or produces actions. Among these capacities one can distinguish between Pavlovian and instrumental conditioning. Pavlovian conditioning [Pav27] consists in the acquisition of responses towards neutral stimuli that have been paired with rewards, such as when one salivates at the bell of the ice cream truck. Instrumental conditioning [Ski38] consists in the expression of a behaviour in order to achieve a goal, such as when one learns to dial a specific number on a phone to call someone. Combined together these mechanisms are at the heart of our learning capacities and their study significantly benefits from computational models.

Reinforcement learning (RL) [SB98], in short learning by trials and errors to decide which action to take in a given situation to achieve a specific goal, is one of the major frameworks used in the current computational models of Pavlovian and instrumental conditioning. As an example of its deep contribution in the expansion of Pavlovian conditioning, the learning algorithm *TD-Learning* [SB81], first developed to explain the prediction capacities of animals in some experimental task, was subsequently shown to rely on a signal that could be paralleled with the activity of some dopaminergic neurons during the experimental task [Sch+97]. Hence, *TD-Learning* successfully linked the expression of a behaviour with some possible underlying neural correlates. With the accumulation of evidences, it is now well accepted that conditioning results from the combination of some kind of reinforcement learning processes.

Surprisingly, while early used in the investigation of Pavlovian conditioning, the modern RL framework is more suited for the investigation of instrumental conditioning, where actions are indeed required to achieve goals. Recent computational models of instrumental conditioning are often the result of a combination of multiple RL systems [Daw+05; Ker+11; Pez+13]. However, recent computational models of Pavlovian conditioning do not rely much more on this general framework but on more specific architectures [Sch+96; MM88; Cou+04; SM07; Jam+12; Has+10]. This is a problem when one investigates the interactions between both types of conditioning as the combination of the various computational models is often not straightforward nor natural.

1.2 Objectives

In the present thesis, we aim at finding key concepts that could be used in RL computational models to allow the study of Pavlovian conditioning, instrumental conditioning and their interactions. Taking inspiration from a variety of experimental data, our intuition is that combining dual-learning and factored representations may help to explain experimental data yet unaccounted for. Dual learning is a commonly accepted concept in the study of instrumental conditioning while factored representations are a concept neglected in RL algorithms of conditioning but often present in the alternative architectures developed to account for Pavlovian conditioning. Especially, we investigated some experimental data about behavioural inter-individual differences in rats undergoing a Pavlovian conditioning task and other experimental data about maladaptive behaviours expressed by pigeons in a supposed interaction task, that could well be explained by such concepts.

1.3 Methods

In order to address these issues, the work presented in this thesis is grounded on a multidisciplinary approach, combining tools or data from neuroscience and artificial intelligence.

On the neuroscience side, we took inspiration from experimental data about maladaptive behaviours and inter-individual variability in conditioning tasks, including behavioural, neurophysiological, neuropharmacological and neuropsychological data. Behavioural analyses involve the observation of animals behaviours in experimental task to investigate their response properties, their capacities, their limits and the strategies developed. Neurophysiology consists in recording the activity of brain regions and/or particular cells, using various techniques such as functional Magnetic Reasonance Imaging (fMRI) or Fast Scan Cycling Voltammetry (FSCV). It helps to investigate the signals on which might rely learning processes that lead to the observed behaviours and locate where values, variables or associations might be stored. Neuropharmacolocy studies the effect of drugs on the nervous system. By injecting drugs that affect specific cells or chemicals, e.g. dopamine, either locally or systemically, it helps to investigate their functions and contributions in the observed behaviours. Similarly, neuropsychology studies the effect of lesions of parts of the brain to identify which and how brain areas contribute to the different aspects of particular behaviours.

On the artificial intelligence side, we mainly use computational models based on machine learning and evolutionary algorithms. Machine learning, from which reinforcement learning algorithms are a subset, are algorithms designed to learn from data in a wide diversity of manners for as many different purposes. In our case, we use it mainly to learn how to successfully accumulate rewards in an efficient way. Evolutionary algorithms are population-based metaheuristic optimization algorithms that can be used to tune algorithms to fit as closely as possible some particular behaviours or results.

In the present work, we first investigated experimental data about conditioning, collected by different approaches, from which we extracted challenging data for the current literature and hints about the mechanisms they might be the result of. Then we developed a computational model with such mechanisms, tuned it with evolutionary algorithms and confronted it to the data for validation.

1.4 Organization

This thesis comprises 2 background chapters, 3 main results chapters and a concluding chapter.

Chapter 2 is an overview of the reinforcement learning framework on which are based the computational modelling aspects of our work. This background provides the necessary notions of this field that are commonly used in computational models of animal conditioning, among which the new computational model developed in the main chapters.

Chapter 3 is an overview of animal conditioning, defining what are Pavlovian and instrumental conditioning and how their interactions are currently understood. It especially lists key phenomena and computational models that fuelled our thoughts regarding the mechanisms expected in the developed computational model.

The 3 main chapters are articles that have been published or are under review in peer-reviewed journals. Chapter 5 introduces a new computational model of animal conditioning that embeds factored representations in reinforcement learning. It uses the model to explain recent experimental data about inter-individual variability in a Pavlovian conditioning task that are conflictual with the current literature. Chapter 6 extends the first one with detailed predictions drawn from the model through new simulations in proposed variants of the original experimental protocol. Chapter 7 shows some generalization abilities of the computational model by applying it to another set of experimental data suggesting inter-individual variability in a different conditioning task about maladaptive behaviours. Each chapter begins with a short introduction that outlines the content of the article and how it is related to the present work.

Finally, Chapter 8 details our contributions, their limits, discusses our architecture choices, and give possible directions for future research.

Chapter 2

Reinforcement learning

2.1 Introduction

Reinforcement learning (RL) is the synthesis of two separate lines of research in different fields, psychology of animal learning (the study of predispositions of animals to act and behave in a certain way, especially in the presence of rewards) [Tho11] and optimal control (solving tasks in an optimal way) [Bel57] (see [SB98] for a deeper review of the steps that led to modern RL). The former provided most of the principles used in current algorithms while the latter especially provided the formalism. As a result, Modern RL offers a normative framework to solve decision making problems in initially unknown environments, achieving goals (e.g. maximizing accumulation of rewards) by trial-and-error searches [SB98].

In this chapter, we first introduce the classical framework of Markov Decision Processes (Section 2.2) used to represent decision problems, its basic notations and key concepts. We subsequently describe the key concepts of the reinforcement learning paradigm (Section 2.3), especially with examples of Model-Free and Model-Based algorithms, and finally action selection processes. Finally we discuss some of the recent advances in reinforcement learning (Section 2.4) that contributed or could contribute to the investigation of animal conditioning.

This chapter does not intend to be exhaustive and focuses specifically on work related to the current thesis. For a deeper overview on reinforcement learning, we refer the reader to dedicated books and reviews [SB98; Kae+96; Sze10; WO12]. For an overview of reinforcement learning in the field of neuroscience, we refer the reader to more specific reviews [Bal+08; Niv09; DB12a]. The subsequent sections are mainly drawn on these studies.

2.2 Markov Decision Processes

Markov Decision Processes (MDPs) provide a mathematical framework (Section 2.2.1) to formalize (Section 2.2.2) and solve (Section 2.2.3) decision problems in uncertain environments [Bel57].

2.2.1 Classical definition

An MDP is formally defined by a tuple $\langle S, A, T, R\rangle$ where

- S is a finite set of states,
- A is a finite set of actions,
- $T : S \times A \times S \to [0, 1]$ is a transition probability function that defines the probability $P(s'|s, a)$ to reach a state s' by doing action a in state s,
- $R : S \times A \to \mathbb{R}$ is a reward function that defines the reward $R(s, a)$ of doing action a in state s.

This formalism implies that the decision problem studied must comply with the *Markov Property*: the future state depends only on the present state and the action taken but not on its past. More formally, it states that

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = P(s_{t+1}|s_t, a_t)$$

where s_t is the state at time t and a_t is the action realized by the agent at time t. For simplification, we usually write s' for s_{t+1}, s for s_t, a' for a_{t+1} and a for a_t.

It also implies that the environment is stable, that is T and R are not supposed to evolve over time. In the current form, all state transitions are supposed to take a same amount of time. It is often the case that they are used in neuroscience to describe experiments which are only stable by blocks or with gradual shifts in probabilities over time. We will come back later to this point and see which limits it implies.

Finally, an MDP is defined as episodic if it includes at least one state which terminates the current episode. It is usually used to define decision problems with unique goals, or repetitive tasks that can be split into independent trials, given that they are clearly delimited. This is usually the case of neuroscience experimental tasks where animals are required to repeatedly achieve the same task or go through the same process for multiple sessions of multiple trials, each trial being separated by some informative cues or resting period, often called the inter-trial interval (ITI).

MDPs are classically represented by directed graphs with states as nodes and transitions probabilities and associated rewards as labelled edges (Figure 2.1). Note that informations provided within states are only here to help the reader but are usually hidden from classical RL algorithms (but see section 2.4).

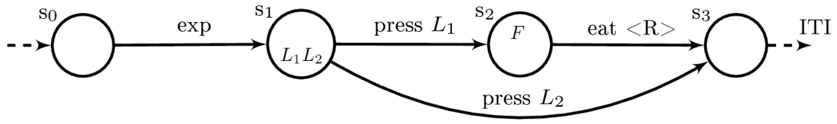


Figure 2.1: Example of an MDP representation. *This MDP could be used to represent a classical instrumental setup. An animal starts in an empty Skinner box (s_0) where at some point 2 levers appear (s_1). The animal must then choose between pressing one lever or the other, the first one leads to the delivery of food (s_2) while the other directly brings the animal back to the terminal state (s_3) which symbolizes the start of the inter-trial interval before next lever appearance. Informations within states are provided for easier readability.*

2.2.2 Value function and policy

Once a decision problem is defined as an MDP, one can define a solution through a *policy* $\pi : S \rightarrow A$ that specifies for each state of the MDP what action should the agent take.

Given a policy π, for each state s one can define a *value function* $V_\pi(s) : S \rightarrow \mathbb{R}$ that describes the sum of cumulative rewards that can be expected by the agent starting from state s and subsequently following policy π. More formally

$$V_\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) | s_0 = s\right] \tag{2.1}$$

where the *discount factor* $0 \leq \gamma < 1$ defines the preference of obtaining a reward immediately rather than delayed (e.g. to get a milkshake today rather than in ten days). This function can be recursively defined as

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V\pi(s'). \tag{2.2}$$

The interest of the current formalism is to find an optimal policy π^* that maximizes the cumulative rewards of the agent over time, that is $\forall\pi\forall s \in S : V_{\pi^*}(s) \geq V_\pi(s)$. Note that while there can be multiple optimal policies, they all share a unique optimal value function

$$V^*(s) = \max_\pi \left[V_\pi(s)\right]. \tag{2.3}$$

Hence, provided Equation 2.2, V^* can be defined as

$$V^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right]. \tag{2.4}$$

The *action-value function* $Q_\pi : S \times A \to \mathbb{R}$ is usually preferred to the value function V_π. It corresponds to the sum of cumulative rewards that can be expected by the agent from taking action a in state s and subsequently follow the policy π:

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_\pi(s'). \tag{2.5}$$

Handling an action-value function Q rather than a value function V permits a comparison between the different actions available in a given state, which is especially useful for action selection algorithms.

In a similar way the optimal action-value function Q^* can be defined as

$$Q^*(s,a) = \max_\pi \left[Q_\pi(s,a) \right] \tag{2.6}$$

or recursively as in Equation 2.4 by

$$Q^*(s,a) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right]. \tag{2.7}$$

Hence, provided with the optimal value function V^* or the optimal action-value function Q^*, an optimal policy π^* can be defined as

$$\pi^*(s) = \operatorname*{argmax}_a \left[Q^*(s,a) \right]. \tag{2.8}$$

To summarize, MDPs are a way of describing a task into a series of experienced states, actions and rewards. From an MDP, an agent can determine an optimal solution, that is which action it should perform in each state in order to maximize the cumulative sum of rewards on the long term.

The optimal value function, action-value function or policy can be found in different ways depending on the amount of information which is a priori given to the algorithm. If the agent has a perfect knowledge of the MDP (he is given a model of the environment which enables it to know exactly what are the consequences of doing each possible action in each possible state), the algorithm can simply consist in propagating the reward information throughout the model, i.e. the directed graph, in order to find the most valuable path (often the shortest to reward). These methods are called dynamic programming. If the agent does not have such an initial model of the world, he can either learn it through exploration and then use it, or directly try to estimate the optimal action-value function and an associated policy. These methods are called reinforcement learning.

2.2.3 Dynamic programming

Given the full knowledge of an MDP, it can be solved by dynamic programming algorithms [Bel57]. They usually define two steps, (1) policy evaluation and (2) policy improvement, which can be combined and repeated of different manners.

Policy evaluation consists in computing $V_\pi(s)$ for all $s \in S$ provided a policy π. It can be incrementally computed given Equation 2.2 by

$$V_\pi^{k+1}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V_\pi^k(s'). \tag{2.9}$$

When $k \to \infty$, V_π^k should converge to V_π [Put95]. Hence, provided a policy π, one can estimate for each state what rewards might be expected by following it.

Policy improvement consists in computing a new and better policy π' given the value V_π of the current policy π. To test and find if there exists such a better policy, one can simply test that for each state s, one of the available actions a can lead to a better state (i.e. $V_\pi(s')$) than the current action proposed by the policy $\pi(s)$, that is

$$\pi'(s) = \operatorname*{argmax}_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_\pi(s') \right]. \tag{2.10}$$

When no better policy can be found, then π' is optimal with regard to the current value function.

Standard dynamic programming algorithms are *Value Iteration* and *Policy Iteration.*

Policy Iteration

In Policy Iteration (Algorithm 2.1) [How60], given an initial policy π_0, we can compute V_{π_0} (step 1) that can be used to yield an equal or better policy π_1 (step 2) which can be again used to compute V_{π_1} and so on, until the best policy is found (i.e. V_{π_t} is optimal). The algorithm is guaranteed to converge to the optimal policy in a finite number of iterations. It is however usually the case that it stops as soon as a near-convergence criterion is reached. Figure 2.2 illustrates such process.

$$\begin{tikzcd}
\pi_0 \arrow[r, "\text{PE}"'] & V_{\pi_0} \arrow[r, "\text{PI}"] & \pi_1 \arrow[r, "\text{PE}"'] & V_{\pi_1} \arrow[r, "\text{PI}"] & \dots \arrow[r, "\text{PE}"'] & \pi^* \arrow[r, "\text{PI}"] & V^* \arrow[l, bend left, "\text{PE}"]
\end{tikzcd}$$

Figure 2.2: Illustration of Policy Iteration. *The algorithm alternates phases of Policy Evaluation (PE) and Policy Improvement (PI).*

Value Iteration

To evaluate a policy π at each time step is costly, as it requires each time to iterate a possibly significant number of times to converge to V_π. Hence, Policy Iteration can be a

Algorithm 2.1: Policy Iteration (from [SB98])

```
  input  : $\emptyset$
  output: $V^*(s), \pi^*$
  Initialize $V_\pi$ and $\pi$ arbitrarily
1 Policy Evaluation
  repeat
      (a) $\Delta \leftarrow 0$
      (b) forall $s \in S$ do
          i. $v \leftarrow V\pi(s)$
          ii. $V_\pi(s) \leftarrow R_\pi(s) + \gamma \sum_{s'} P(s'|s, \pi(s))v(s')$
          iii. $\Delta \leftarrow \max(\Delta, |v - V_\pi(s)|)$
  until $\Delta < \varepsilon$ (with a positive small $\varepsilon$)
2 Policy Improvement
  stable $\leftarrow$ true
  forall $s \in S$ do
      (a) $b \leftarrow \pi(s)$
      (b) $\pi(s) \leftarrow \mathrm{argmax}_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')\right]$
      (c) if $b \neq \pi(s)$ then stable $\leftarrow$ $false$
  if not stable then go to step 1
  return $V_\pi$ and $\pi$
```

Algorithm 2.2: Value Iteration (from [SB98])

```
  input  : $\emptyset$
  output: $V^*(s), \pi^*$
  Initialize $V_\pi$
  repeat
      (a) $\Delta \leftarrow 0$
      (b) forall $s \in S$ do
          i. $v \leftarrow V_\pi(s)$
          ii. $V_\pi(s) \leftarrow \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s')\right]$
          iii. $\Delta \leftarrow \max(\Delta, |v - V_\pi(s)|)$
  until $\Delta < \varepsilon$ (with a positive small $\varepsilon$)
  $\pi^*(s) \leftarrow \mathrm{argmax}_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)V_\pi(s')\right]$
  return $V_\pi$ and $\pi^*$
```

time consuming algorithm. Value Iteration [Bel57] offers a nice alternative. It has been shown that for some problems, it is not necessary to wait until the converge of V_π and a small number of iterations in the policy evaluation step are sufficient for the Policy Iteration algorithm to converge. More precisely, it is possible to only iterate once, which gives Value Iteration (Algorithm 2.2). Note that the two steps can easily be merged by using the max operator in the second step (ii) of the policy evaluation process.

2.3 Reinforcement learning

Dynamic programming algorithms are efficient methods when the problem is fully known. However, it is often the case, especially in neuroscience experiments, that agents do not have a complete knowledge of their environment and need to interact and move within it to acquire useful information for eventually solving the problem.

The computational field of reinforcement learning [SB98] addresses such limitation. The agent starts with no prior knowledge about the environment and learns from the consequences of its actions by trial-and-error experiences. It learns the optimal value and/or an optimal policy over time. Interestingly, while based on the MDP formalism, such methods are well suited for changing environments as they can revise their beliefs about the world's dynamic over time. This is a useful property in animal conditioning experiments.

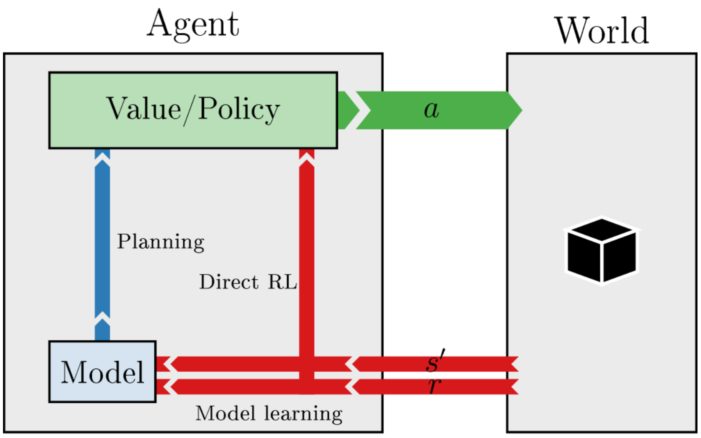


Figure 2.3: Illustration of reinforcement learning algorithms. *The agent has no knowledge of the world's dynamic (black box) and acquires it through trial-and-error interactions, taking an action a and observing the resulting new state s' and possible reward r. Model-Based algorithms learn a model (Model learning) of the world from observations from which they infer a value function/policy (Planning). Model-Free algorithms directly learn a value function/policy from observations (Direct RL). These values/policy are used by action selection mechanisms to select the next action.*

There are two main categories of RL algorithms, using different pathways to achieve the same goal (Figure 2.3). Model-Based algorithms (Section 2.3.2) incrementally learn an internal model of the world by experience from which they can infer values over states that help to guide behaviour. Model-Free algorithms (Section 2.3.1) directly learn these values by experience without relying on some internal model. While these algorithms are usually seen as the full process of learning, planning and acting, they mainly define the process of the two first steps, which can then be combined with various action selection mechanisms (Section 2.3.3) that only need values over state-action pairs to work.

2.3.1 Model-Free algorithms

In the following section, we present 3 Model-Free (MF) algorithms (Actor-Critic, Q-Learning and SARSA) that have been extensively linked to instrumental and Pavlovian phenomena. MF algorithms have in common that they rely on the incremental learning of value functions by trial-and-error experiences without the help of an internal model of the world. These algorithms derive from the Temporal Difference Learning principle.

Temporal Difference Learning

The Temporal Difference Learning principle (TD-Learning) [SB87; Sut88] offers a way to estimate the value function over time through experience with the environment. It does not require the knowledge of T or R nor builds a representation of it. It essentially relies on the key concept of Reward Prediction Error (RPE).

The RPE signal is the difference between the current estimation by the agent of the value function $\hat{V}_t(s_t)$, i.e. its expectation, and the value of its last observation $r_t + \gamma \hat{V}_t(s_{t+1})$. Provided the recursive definition of the value function (Equation 2.2) and that $R(s, a)$ and $P(s'|s, a)$ can be approximated by the last observation $\langle r_t, s_{t+1} \rangle$, there should be no difference (i.e. a null RPE) if the value function $\hat{V}_t$ is correct. This signal is formally defined as

$$\delta_t \leftarrow \underbrace{r_t + \gamma \hat{V}_t(s_{t+1})}_{V(\text{observation})} - \underbrace{\hat{V}_t(s_t)}_{V(\text{expectation})} \tag{2.11}$$

where r_t is the reward retrieved after doing action a_t in state s_t and ending in the new state s_{t+1}.

Hence, if the signal is not null, it implies that the current estimate of the value function needs to be revised, which is done by

$$\hat{V}_{t+1}(s_t) \leftarrow \hat{V}_t(s_t) + \alpha \delta_t \tag{2.12}$$

where the learning rate $0 \leq \alpha \leq 1$ defines at which rate to revise the estimated value function, i.e. how much the agent relies on the last observation versus its cumulative history. While it is common to decrease α over time in a stable environment, such that $\hat{V}_t$ eventually converges to the correct value V^* on an infinite horizon, a fixed α allows the agent to adapt in changing environments, which is often a need in computational neuroscience tasks.

As presented in more details in Section 3.3.2, the RPE signal used in TD-learning has been shown to mirror the phasic activity of some dopamine neurons, increasing at the delivery of a better than expected reward, diminishing at the omission of an expected reward and not responding when the delivered reward could be expected. This discovery greatly influenced the computational contribution of the RL framework to the neuroscience field of conditioning [Sch+97; Sch98; Sch10; Sch13; Gli11; Har+14; BG05].

MF algorithms that rely on the TD-Learning principle have the advantage of being fast to execute, as they almost require a single computationally simple update at each agent step. The associated drawback results in relatively slow learning and revision capacities.

The TD-Learning principle provides an efficient tool to maintain and update a valued estimation of all the situations that an agent can face. However, without knowledge of the world transition function T, the agent has no way to decide which action will guide him to the most rewarding situations when only provided with a value function V. The subsequent algorithms offer a solution to this problem by providing values over state-action pairs, which can be subsequently used by action selection mechanisms.

Actor-Critic

Actor-Critic methods [Bar+83] are split into two components. The *Critic* estimates the value function V of the current policy of the agent over time, and is a direct application of the TD-Learning principle. The *Actor* maintains a function p that defines the agent's *preference for* each action in each situation and uses the RPE signal computed in the *Critic* to revise it (Figure 2.4).

Such methods are of particular interest to neuroscientists because a parallel can be made between their architecture and the anatomy of the Basal Ganglia. Its ventral part seems to estimate and store values about expected rewards [CS03; Kha+08] and directly projects on dopaminergic neurons, while its dorsal part seems to learn the value of actions [Sam+05].

When the action taken at time t results in a positive signal δ_t, it implies that the action taken improved the prospects for future rewards. Therefore, such an action should be selected more often. The inverse is also true for a negative signal. Hence, the preference function p can be incrementally revised by

$$p(a_t|s_t)_{t+1} \leftarrow p(a_t|s_t)_t + \eta\delta_t \tag{2.13}$$

where $0 \leq \eta \leq 1$ is another learning rate. Hence, for a given state s, the estimated best action is the most valued one in p, that is $\max_a p(s,a)$.

SARSA

SARSA (for State-Action-Reward-State-Action) is built on a straightforward revision of the TD-Learning principle to embed actions by replacing V by Q in Equations 2.11 and 2.12, in the following way

$$\hat{Q}_{t+1}(s_t,a_t) \leftarrow \hat{Q}_t(s_t,a_t) + \alpha \underbrace{\left[r_t + \gamma\hat{Q}_t(s_{t+1},a_{t+1}) - \hat{Q}_t(s_t,a_t)\right]}_{\delta} \tag{2.14}$$

where the next action a_{t+1} has already been chosen by the system.

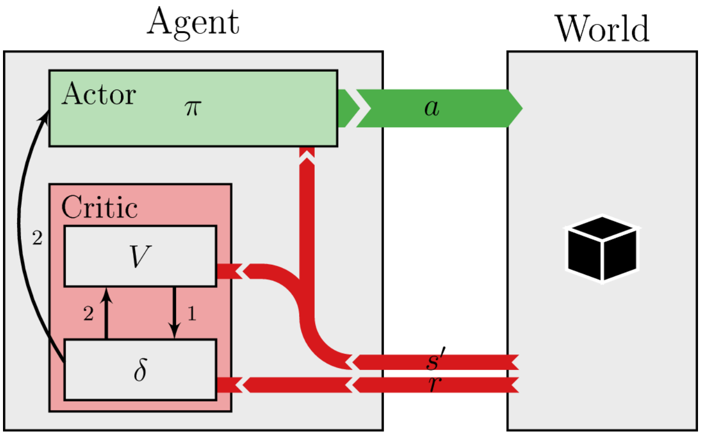


Figure 2.4: Illustration of the Actor-Critic architecture. *The agent is composed of two modules: an Actor and a Critic. The Critic learns and updates a value function V with the temporal learning principle, i.e. using a reward prediction error signal δ ②. This signal is also used by the Actor to build a policy π ②. This signal is computed as the difference between the value of what was expected from doing action a in the current state and what was actually observed $\langle s', r \rangle$ ①.*

This method is defined as *on-policy* as its learning phase, the RPE signal, depends on the actions a_t and especially a_{t+1} chosen by the agent policy.

Q-Learning

Q-Learning [WD92] is an *off-policy* alternative to SARSA, where the RPE signal does not rely on the agent policy, i.e. it does not need the next chosen action to be computed. Furthermore, the updated value can be taken into account in the choice of the next action. Equation 2.14 is revised as

$$\hat{Q}_{t+1}(s_t, a_t) \leftarrow \hat{Q}_t(s_t, a_t) + \alpha \underbrace{\left[r_t + \gamma \max_{a'} \hat{Q}_t(s_{t+1}, a') - \hat{Q}_t(s_t, a_t) \right]}_{\delta} \tag{2.15}$$

where the max operator suggests that the RPE is computed with respect to what is believed to be the best action on the subsequent state. It is not necessarily the case that such action will be chosen.

Interestingly, while Q-Learning and SARSA are proven to converge to the true optimal policies over time [WD92; Sin+00], their RPE signals can present very distinct patterns in specifically designed tasks. Based on the hypothesis that phasic dopaminergic activity encodes RPE signals, different experiments were conducted to determine whether such patterns would suggest Q-Learning like or SARSA like RPEs [Mor+06; Roe+07]. Conflicting results make this question still under investigation [Bel+12a; Bel+13; Bel+12b] (see Section 3.3.2).

2.3.2 Model-Based algorithms

Model-Based algorithms (MB) lie between Dynamic programming approaches, which assume a complete knowledge of the world and can infer an optimal policy without the need to interact with it, and MF algorithms that avoid the need of such knowledge but require direct and repeated interactions.

Such algorithms rely on the acquisition by experience of some knowledge regarding the structure of the world and use it to infer an optimal policy with regard to their current beliefs.

Dyna-Q

Dyna-Q [Sut90] combines MF updates with a learned internal model of the world (i.e. the transition function T and reward function R) to accelerate and refine the computation of an action-value function.

At each time step, the world is revised with the new observation, either by accumulating samples of observations or building an estimate of the transition function $\hat{T}$ and reward function $\hat{R}$, for example with the following formulae

$$\hat{T}(s,a,s') \leftarrow \begin{cases} (1-\alpha_T) \times \hat{T}(s,a,s'') + \alpha_T & \text{if } s' = s'' \\ (1-\alpha_T) \times \hat{T}(s,a,s'') & \text{otherwise} \end{cases} \tag{2.16}$$

$$\hat{R}(s,a) \leftarrow \hat{R}(s,a) + \alpha_R(r - \hat{R}(s,a)) \tag{2.17}$$

where $0 \leq \alpha_T \leq 1$ and $0 \leq \alpha_R \leq 1$ are two learning rates that represent the speed at which new experiences replace old ones.

At each time step, the agent internally replays multiple time some observations that already have occurred in the past and update the action-value function Q accordingly as in Equations 2.15 or 2.14 (Algorithm 2.3).

One of the drawbacks of Dyna-Q is the requirement of the arbitrary free parameter N (Algorithm 2.3), that fixes a limit in the number of off-line iterations. If close to 0, the agent is almost equivalent to an MF algorithm, requiring more memory usage (to store its internal model), with the same results. A large N, which would successfully make Q converge to Q^* can significantly slow down the update phase, making it unsuitable for real-time environments. Such parameter might be replaced by other convergence criteria (e.g. similar to the one used in dynamic programming methods). Dyna-Q can be revised to improve how states and actions are selected in the simulation step to optimize convergence (e.g. Prioritized sweeping [MA93] or Queue-Dyna [PW93]).

Dyna-Q has been used with success in some neuroscience experiments. For example, it explained well how some subjects implicitly learned a model of the biomechanical costs of their movements during a motor babbling phase and how they implicitly integrated these costs into subsequent choices to reach equally distant targets but which would require moves with different costs [Cos+13].

Algorithm 2.3: Dyna-Q (from [SB98])

```
Initialize $Q$ and $M = \langle T, R \rangle$ $\forall s \in S$ and $\forall a \in A$
repeat
    $s \leftarrow$ current non terminal state
    $a \leftarrow$ select an action
    observe $\langle s', r \rangle$ from taking $a$ in $s$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\,[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
    update $M$ given $\langle s', a \rangle$
    repeat
        $s \leftarrow$ random previously observed state
        $a \leftarrow$ random previously taken action in $s$
        retrieve $\langle s', r \rangle$ from $M$
        $Q(s,a) \leftarrow Q(s,a) + \alpha\,[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
    until $N$ times
until forever
```

Q-Learner / Tree search methods

Provided with a learned internal model (Equations 2.16 and 2.17) one can directly rely on the formal definition of Q (Equation 2.5) without using the TD-Learning principle. One can use dynamic programming methods on the internal model [Glä+10; BT03; KS02] at each time step. This is computationally expensive as the tree of the possible paths can grow exponentially with the number of states. In more realistic situations, (e.g. complex navigation tasks in real-time environments), the high number of states makes such MB systems less efficient than expected [Cal+12; Ren+14].

It is also possible to use lookahead methods or shortest path algorithms to find the full plan of actions (or at least for multiple steps) that leads to a specified reward [KS02; Wal+10; KS06; Bro+12]. One can then use the formal definition of Q (Equation 2.5) in an efficient way to back-propagate values from the goal state to the current state. Such methods are however only tractable for simple MDPs. It is not uncommon that MB processes are described as tree searches in conditioning studies, even if their implementation usually rely on estimations of Q value functions.

In some conditioning experiments, it has been shown that animals seem to take the time to plan ahead before acting (e.g. showing head movements towards alternative paths at crossing [Red+07; Red+08] before any subsequent engagement). Furthermore, if after some training in a maze, the optimal path is blocked, rats directly choose the next shortest path that should lead to reward, which also suggests that they may use a topological representation of their environment to take decisions [Mar+11].

2.3.3 Action selection

Provided with an action-value function Q (or p), there is an infinite number of ways to define a policy π that guides the behaviour of the agent. We present 3 major action selection algorithms.

Greedy

The simplest algorithm consists in a greedy policy that always takes the most valued action, that is

$$\pi_{\text{Greedy}}(s) = \underset{a}{\operatorname{argmax}} \left[Q(s,a) \right]. \tag{2.18}$$

This policy is a wrong solution for reinforcement learning algorithms as it could lead to a local maximum. Indeed, such a policy favours paths that have already been taken and rewarded, and can miss alternative higher rewarding paths, especially in changing environments. This problem is called the exploration/exploitation trade-off.

Exploration/Exploitation trade-off

The exploration/exploitation trade-off is a classical dilemma of reinforcement learning, which is to decide how to balance between exploitation – using the already acquired knowledge to optimize cumulative rewards – and exploration – acquiring new knowledge that could improve the general behaviour. For example, when confronted to a changing environment, it is impossible to know if a path already visited and discarded at first might actually now lead to a high reward. There is no optimal solution to this problem.

ε-Greedy

One solution is to be greedy most of the time, except for some exploration steps drawn with a small probability ε that guarantees sufficient exploration to avoid local minima. More formally

$$\pi_{\varepsilon\text{-Greedy}}(s) = \begin{cases} \text{random action} & \text{with a probability } \varepsilon \\ \pi_{\text{Greedy}}(s) & \text{otherwise} \end{cases} \tag{2.19}$$

where $0 < \varepsilon < 1$.

SoftMax

In ε-Greedy, when an action is of almost equal value as the best action, it will still be selected only with a probability ε. Any action other than the best one, is selected with an equally low probability disregarding its relative value compared to others. However, if an action has a very bad value (especially with negative rewards as in punishment experiments) one would expect to avoid it most of the time relative to, for example, actions with neutral values. If an action is almost equivalent to the top most action, we would expect it to be chosen almost as often.

Based on this idea, the softmax function provides such an action selection mechanism, as it selects actions based on probabilities that are built from the state-action values, which are computed with

$$\pi_{\text{softmax}}(s, a) = \frac{e^{Q(s,a)/\tau}}{\sum_a' e^{Q(s,a')/\tau}} \tag{2.20}$$

where $0 < \tau < \infty$ is called the *temperature*. At high temperatures ($\tau \to \infty$) all actions have almost the same probability to be chosen as Q-values become negligible given τ. At low temperatures ($\tau \to 0^+$), the probability of selecting the action with the highest value tends to 1.

This process is widely used in computational neuroscience [Daw+06b; Day+06; Glä+10; Hum+12; Daw+05; Ker+11; Huy+12; Doy+02; Red+07]. The computational model developed in this work also relies on such mechanism.

2.4 Extensions of the classical framework - Factored representations

Markov Decision Processes have benefited from various extensions to address very different problems, usually leading to new or revised versions of reinforcement learning algorithms. MDPs were extended to continuous times and actions [Bai93; Duf95; Doy00]; for partially observable environments (POMDPs) [Jaa+95; Hau00]; to allow factored representations [Bou+00; Deg+06; VB08]. Another example is the concept of hierarchical RL where problems can be defined at multiple details levels such that one can define sequences of actions as subroutines (options) to be played as one action at a higher level [Sut+99; Bot+09; Bot12; Diu+13]. Most of these extensions have been pushed back into the field of neuroscience and nourished some animal conditioning investigations [Daw+03; Daw+06c; Bot+09; Doy00; RF+11]. However, to our knowledge factored representations for reinforcement learning have been left apart.

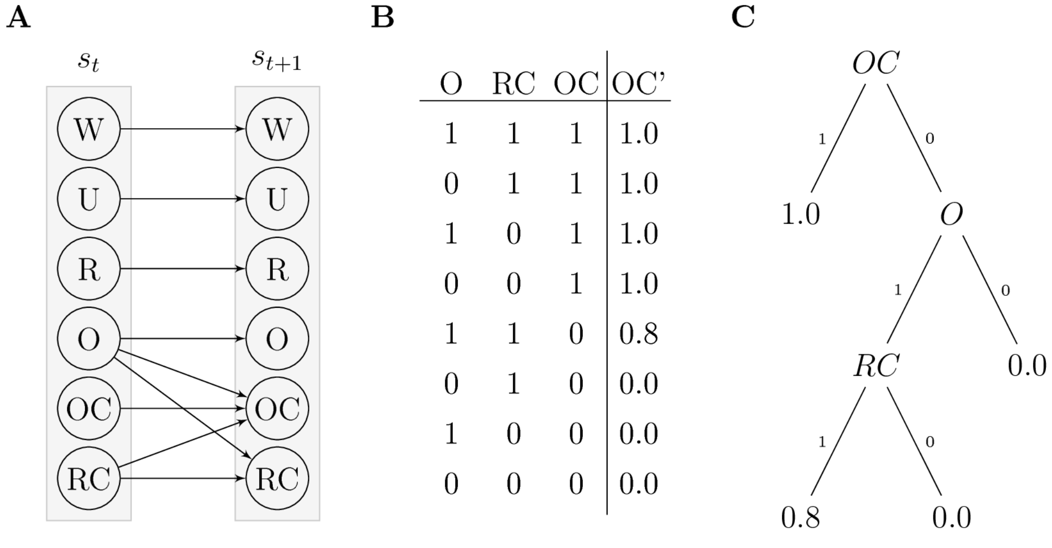
The original algorithms in the literature that use factored representations rely on MB learning principles, while in the present work, we develop an algorithm based on MF learning principles. This part of the manuscript will thus describe the original algorithms in a mainly informational manner, in order to understand on which principles we implemented factored representations for MF learning without sticking to the original formalism.

The idea of factorization comes from the necessity to deal with large-scale problems. The standard MDP representation and classical algorithms do not scale well to high dimensional spaces and ends up requiring too much physical space or computation time, a phenomenon named the *curse of dimensionality* [Bel57]. We illustrate the principle of factorization and associated algorithms through the common *CoffeeRobot* example task [Bou+95; Bou+00], in which, one robot needs to go buy a coffee across a street and deliver it back to an employee, earning extra credits if it does not get wet in case of rain.

Real application problems are often described through a set of parameters that describe the current state of the system. Hence, the set of states S can formally be described through a set of random variables $X = \{X_1, \ldots, X_n\}$ where each variable X_i can take several values. A state is therefore an instantiation of X. It is also commonly the case that the random variables are binary, that is $X_i \in [0, 1]$. In such a case, states can be defined by the active/present variables in the situation they describe. With factored representations, informations embedded within states are explicitly made available to the algorithms. The *CoffeeRobot* task is described by a set of 6 binary variables: the robot is wet W, the robot has an umbrella U, the robot is in the office O (outside otherwise), it is raining R, the robot has the coffee RC or the employee has it EC. Hence, the set $\langle RC, R, O\rangle$ describes the state where the robot has a coffee, is in the office while it is raining, but has not delivered the coffee to the owner yet, and is neither wet nor has an umbrella. While simple at first sight, this toy problem already has $2^6 = 64$ states.

Actions are still described as in the standard MDP framework. For example, the robot can go to the next location (*Go*), buy a coffee (*buyC*), deliver the coffee (*delC*) to the owner and get an umbrella if in the office (*getU*).

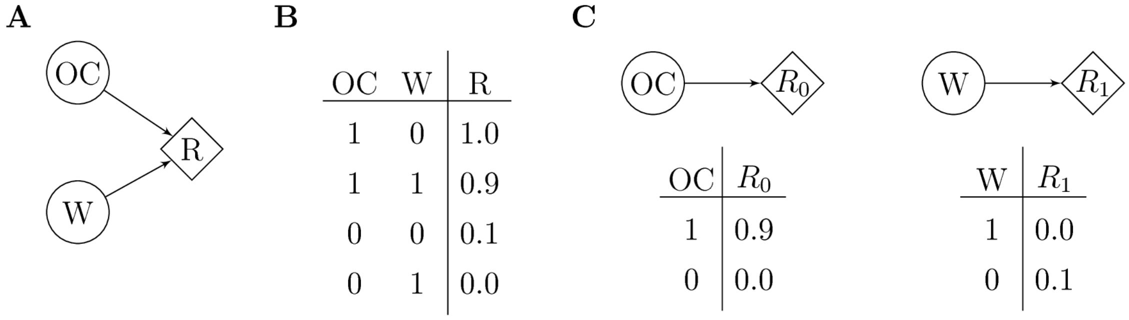
T and R are also usually redefined to take advantage of the factored representation, describing problems through *Factored MDPs* [Bou+95; Bou+00; Deg+06; VB08]. Factored MDPs use Dynamic Bayesian Networks [DK89] to define dependencies between variables, combined with compact conditional probability distribution described through trees. The important idea is that some aspects of the task are independent of others. In our example, the fact that it is raining has no impact on the success of delivering the coffee (Figure 2.5).



O	RC	OC	OC'
1	1	1	1.0
0	1	1	1.0
1	0	1	1.0
0	0	1	1.0
1	1	0	0.8
0	1	0	0.0
1	0	0	0.0
0	0	0	0.0

Figure 2.5: Illustration of the *CoffeeRobot* task state transition. *Representation of the transitions of doing the action of delivering the coffee (*delC*).* **(A)** *Dependencies between variables (DBN). Associated conditional probability distribution $P(OC'|O, RC, OC, delC)$ under tabular form* **(B)** *and compact tree form* **(C)***. The full problem is described by one DBN per action and one conditional probability distribution for each variable and actions. If the owner already has a coffee, it will still have it at the next state (1.0). If the owner does not have a coffee, but the robot is in the office and has a coffee and performs action* delC*, the owner might successfully retrieve it 80% of the time, i.e. failures are possible.*

It is also possible to split the reward function into orthogonal goals that can be dealt with independently (Figure 2.6).



OC	W	R
1	0	1.0
1	1	0.9
0	0	0.1
0	1	0.0

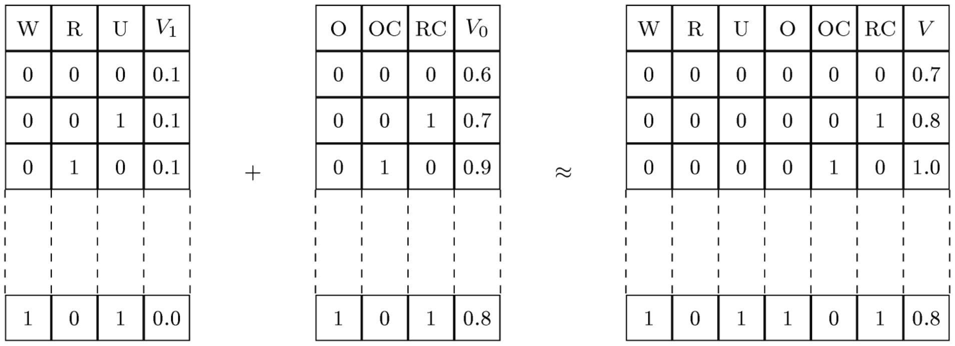
OC	R_0
1	0.9
0	0.0

W	R_1
1	0.0
0	0.1

Figure 2.6: Illustration of the *CoffeeRobot* task reward function. **(A)** *Dynamic Bayesian Network representation of the reward function.* **(B)** *Associated conditional probability distribution $P(R|OC, W)$.* **(C)** *Representation of the function once split into sub-goals. Delivering the coffee and being dry are two rewarding components. However, they are independent of each other and can be dealt separately.*

Value function approximations and *factored reinforcement learning* (FRL) are the main methods that have been developed to take advantage of factored representations.

Value function approximations [Doy+02; Kha+06; Elf+13] attempt to split problems into orthogonal sub-problems making computations easier and providing valuations that can then be aggregated to estimate the value of states. This paradigm is illustrated in Figure 2.7, where one can split the problem according to the two orthogonal goals of the reward function (Figure 2.6).



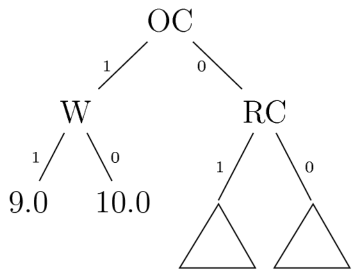
W	R	U	V_1
0	0	0	0.1
0	0	1	0.1
0	1	0	0.1
1	0	1	0.0

O	OC	RC	V_0
0	0	0	0.6
0	0	1	0.7
0	1	0	0.9
1	0	1	0.8

W	R	U	O	OC	RC	V
0	0	0	0	0	0	0.7
0	0	0	0	0	1	0.8
0	0	0	0	1	0	1.0
1	0	1	1	0	1	0.8

Figure 2.7: Illustration of value function approximations methods. *The problem is separable into 2 different sub-problems (1) being wet, (2) delivering a coffee, such that the aggregation of values over their respective sub-spaces defines the value function of the global problem. Note that the real problem might be more complicated and the sub-problems have overlapping domains.*

Factored reinforcement learning [Bou+00; Deg+06; VB08] reduces the physical space needed to store the value function by representing it with a tree which leaves, i.e. the values, can account for multiple states. For example, if the owner already has a coffee and the robot is not wet, all other variables are irrelevant to decide the value of the state (Figure 2.8). Planning and learning algorithms also rely on Factored Markov Decision Processes to optimize computations. For example, one can take advantage of the tree form of a conditional probability distribution to compute a Q-value for a particular action for all states at once (e.g. Structured Value Iteration [Bou+00]). Figure 2.9 illustrates such a possibility on the simplified version of the *CoffeeRobot* problem where only variables U, R and W are taken into consideration, the robot is only interested in not getting wet.



$\approx$

W	R	U	O	OC	RC	V
0	0	0	0	0	0	0.7
1	·	·	·	1	·	0.9
0	·	·	·	1	·	1.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	1	1	0	1	0.8

Figure 2.8: Illustration of model-based factored reinforcement learning value functions. *All variables are not meaningful at all time to value a state. In the* CoffeeRobot *example, if $OC = 1$ in the current state, only W is relevant to determine the value of the state.*

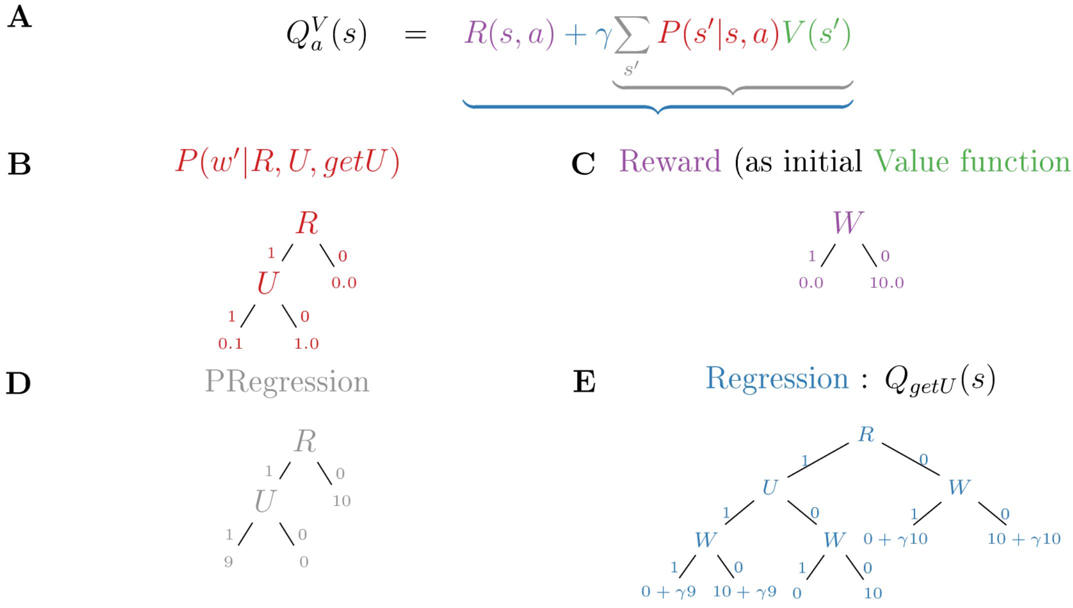


Figure 2.9: Illustration of Structured Value Iteration. *Example of the computation of a Q value for all states for action getU* **(A)** *optimized by the use of trees. Starting from the conditional probability distribution $P(w'|R,U,getU)$* **(B)** *and the reward function R as the current value function V* **(C)**, *we apply a PRegression step* **(D)**, *which computes the gray underlined part in the formula, and then a Regression step* **(E)**, *which computes the Q value.*

In the main chapters, we introduce a version of an MF algorithm that also takes advantage of factored representations. However, in contrast to FRL algorithms, we do not intend to build a compact value function nor to infer the value of states from values of features but rather make these values compete in the choice for the next action.

Chapter 3

Animal conditioning

3.1 Introduction

The study of animal conditioning can be broadly divided between Pavlovian conditioning [Pav27] and instrumental conditioning [Ski38]. In instrumental or operant conditioning (Section 3.2), animals are required to make actions (e.g. pulling a chain or refraining from it) for rewards to be delivered or punishments to be avoided. The study of instrumental conditioning aims at investigating the strategies developed by animals to achieve such tasks and understanding their underlying mechanisms. In contrast, in Pavlovian or classical conditioning (Section 3.3), neutral stimuli (e.g. sound or light) are paired with rewards or punishments and delivered regardless of the animal's behaviour. With training, animals usually develop responses (e.g. salivation or freezing) towards these neutral stimuli. The study of Pavlovian conditioning aims at measuring these behaviours and understanding their underlying mechanisms.

In this chapter, we first develop instrumental conditioning (Section 3.2) and how associated computational models mainly rely on the reinforcement learning framework. Then we develop Pavlovian conditioning (Section 3.3), with an emphasis on some phenomena that are challenging to classical RL algorithms and classical MDPs, and give a brief overview of the wide variety of associated computational models. Finally, we discuss phenomena that combine Pavlovian and instrumental aspects (Section 3.4) and their sparse related computational models.

3.2 Instrumental conditioning

Instrumental conditioning results from the confrontation of animals to tasks that actively require specific sequences of actions (often a single action) to maximize rewards or avoid punishments [Ski38; Tol38; Tol39]. Such experiments allow the study of the strategies developed by animals to construct or revise such sequences, especially when confronted with stochastic results or changing environments. Such tasks usually introduce some operant objects (e.g. a lever or a chain) which, when actively engaged (e.g. pressed or pulled), lead to the next phase of the task up to reward delivery or its omission. An example is provided in Figure 3.1.

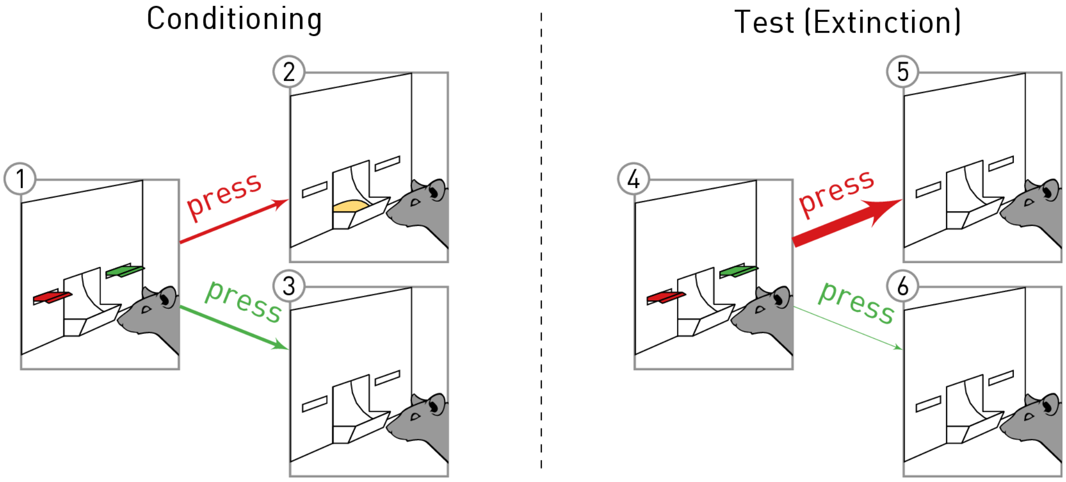


Figure 3.1: Example of an instrumental conditioning task. *A rat is presented with two levers ①. When one press occurs, levers are retracted. If the correct lever is pressed ②, food is delivered in a magazine. If the wrong lever is pressed ③, reward is omitted. Conditioning is assessed in a second phase, where the rat is presented with the two levers in extinction ④(no reward is available). If successfully conditioned, the rat shows a preference for the previously rewarded lever ⑤.*

While not shown in Figure 3.1, it is often the case that a single action (e.g. pressing a lever) should not be sufficient to get a reward. Hence, experimental tasks usually involve different reinforcement schedules [Dom14; Nev+01; BB08]. Rewards may be delivered after a number of responses (ratio schedules) or after some delay interval (interval schedules). The number of responses required might be fixed (FR) or vary during the task (VR). The interval can also be fixed (FI) or vary during the task (VI). One can also distinguish between variable and random schedules, the latter being a subset of the former that arranges a constant probability across these numbers of responses (RR) or delay interval (RI) involved. The type of schedule may influence the expressed behaviours [Nev+01; Wil+12; Bau93]. Such reinforcement schedules aim first at avoiding any confusion between an engagement resulting from conditioning versus a random interaction.

It is commonly assumed that animals do not always rely on the same strategies to guide their behaviour and that one can distinguish between Habitual and Goal-Directed behaviours [AD81; Ada82]. In the following, we discuss this dichotomy and experiments allowing to distinguish it (Section 3.2.1), theories on their expression (Section 3.2.2), and computational models that have been developed to account for it (Section 3.2.3).

For a big picture about instrumental conditioning, especially regarding biological links, we refer the reader to dedicated reviews [DD13; Yin+08; BO09; Dom14].

3.2.1 Habitual versus Goal-Directed behaviours

It is now well accepted that animals' behaviours rely on two learning processes when involved in instrumental tasks, one Goal-Directed (GD) and the other Habitual [AD81; Ada82; Fan+13; Ash+10; Tho+10; KH12; Bal+07; BO09; Bro+11]. A behaviour is considered Goal-Directed if it clearly (1) links actions to their consequences and (2) is guided by a desirable outcome, such that it quickly adapts to changing situations or evolutions in its motivational state. A behaviour is considered Habitual when it does not respect the preceding conditions, i.e. it is decorrelated from the expected result of actions, in their consequences or resulting outcomes.

In simple and stable instrumental tasks, Goal-Directed and Habitual behaviours cannot be distinguished, as they produce similar undistinguishable outputs. It has been shown that behaviour usually shifts from Goal-Directed to Habitual with overtraining (extensive training on the same task, on multiple days and multiple trials per days) [AD81]. This property has been deeply investigated through *outcome devaluation* and *contingency degradation*, two processes (implemented in many different ways) that help to distinguish between both phenomena.

Outcome devaluation

The outcome devaluation procedure consists in reducing the value of the instrumental task outcome and observing its effect on the behaviour [AD81; Ada82; DB94; Dic+95; Tri+09; Val+07; CR85; BD98b]. This devaluation is usually done by specific satiation – giving free access to the current reward to animals – or with paired illness – combining the current reward with some nauseous substance. After an initial phase of instrumental conditioning, animals undergo a devaluation protocol and are subsequently tested in extinction. If animals produce less instrumental responding for an outcome that has been devalued than for an outcome that has not, then the behaviour of these animals is devaluation sensitive and is assumed to be driven by the outcome, i.e. it is Goal-Directed. Figure 3.2 shows an example of such a devaluation procedure. If animals persist to respond with the same intensity as before devaluation, their behaviour is disconnected from the expected outcome value, i.e. Habitual.

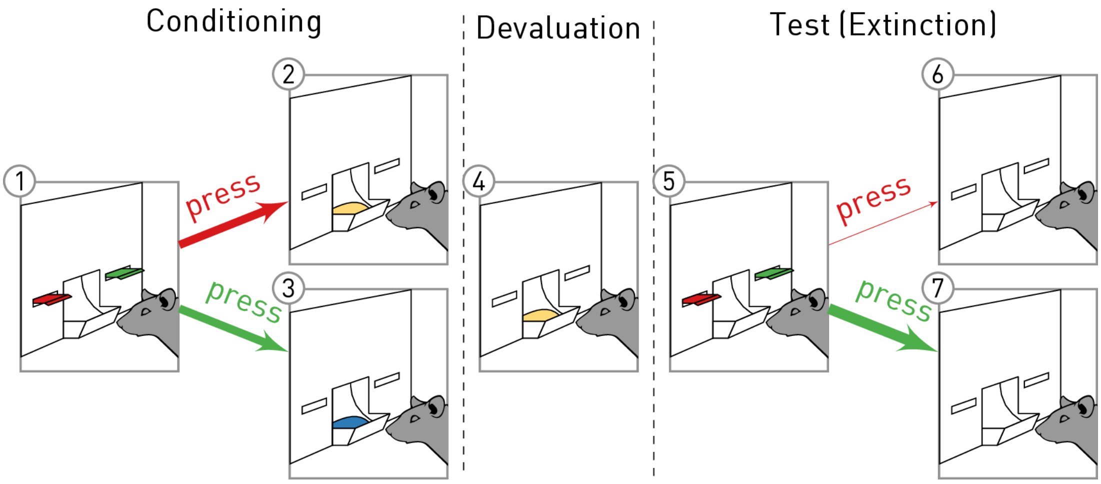


Figure 3.2: Illustration of devaluation procedure. *After a first phase of instrumental conditioning (e.g. with 2 levers) ①②③, the reward of one lever is devalued ④(e.g. by satiation or illness pairing). In a third test phase, the persistence or suppression of conditioning is assessed in extinction ⑤⑥⑦. If the behaviour is sensitive to the devaluation, during the extinction test the animal chooses more often to press the non-devalued lever ⑦.*

Contingency degradation

The contingency degradation procedure consists in breaking the current consequence (resulting state or reward) of an action in the sequence that initially led to the reward, or making such contingency weaker, for example by delivering rewards without the need for acting [KC03; BD98a; CR86; DM89]. The simplest case is to stop the delivery of the reward in the presence of one action. As in outcome devaluation, if the animal reduces its level of instrumental responding in extinction, its behaviour is driven by the outcome, i.e. Goal-Directed. If the animal persists to behave at a similar level as before contingency degradation, the behaviour is disconnected from the outcome expectancy, i.e. Habitual.

3.2.2 Behavioural expression and Neuronal correlates

While the transition from Goal-Directed towards Habitual behaviours with overtraining is a standard phenomenon of the literature [AD81; Ada82; DB94; Dic+95; Tri+09; Val+07; CR85; BD98b; KC03; BD98a; CR86; DM89], other experiments showed that overtraining is not decisive to define which process is currently driving the behaviour [KD10; YK06]. At the behavioural level, it has been shown that stress seems to induce Habitual behaviours [SW09; SW11], as limited working-memory capacities [Ott+13] and distractions [Foe+06]. The characteristics of the responses (pulling a chain versus pressing a lever) have also been suggested to elicit different behaviours [Fau+05].

At the biological level, a significant number of studies were realized on the importance of the Dorsal Striatum (DS) for instrumental conditioning [BO09]. Using lesions, pharmacological interventions or brain recordings, studies have shown correlates of Goal-Directed behaviours in Dorsomedial Striatum (DMS) [Boo+09; Wit+09; Glä+10; KC03; Wun+12; Yin+05; Ske+14; Fan+13]. In contrast, the expression of Habitual behaviours was correlated with the Dorsolateral Striatum (DLS) [Wun+12; Yin+04; Yin+06; YK06; Fan+13; Ske+14]. Lesions studies in the prefrontal cortex could induce a switch between systems [OB05; KC03; Smi+13; MC01; BB00; Rag+02]. Furthermore, Goal-Directed and Habitual capacities have been shown to differ across individuals [Ska+13].

Hence, while Habitual and Goal-Directed behaviours have clearly been identified and shown to rely, at least partially, on different brain regions, how one system comes to control the behaviour is still unclear. The current literature suggests a potentially complex integration or competition mechanism.

3.2.3 Theories and Models

The reinforcement learning framework has been decisive in the computational modelling of instrumental phenomena [DD13; DB12a; Huy+12; Daw+05; Ker+11; Bei+11; Dol+12; Daw+11; SD12], the core of which being the distinction between Model-Based (MB) and Model-Free (MF) algorithms.

It is now commonly accepted that MB reinforcement learning is well suited to explain Goal-Directed behaviours [Daw+05; DD13; BO09; Dol+12]. As it relies on an internal model of the world (transitions and reward values) to build and search for the best plan of actions, it can immediately adapt if the dynamic of the world is unexpectedly modified, such as in outcome devaluation or contingency degradation. Moreover, maintaining an internal model of the world and inferring on it *on the fly* is costly in term of computation time and requires some space capacity. This is consistent with observations that Goal-Directed behaviours require sufficient attention and working memory capacities [Ott+13; Foe+06; Ger+14].

MF learning on the other side, is well suited to explain habitual behaviours [Daw+05; DD13; BO09; Dol+12]. Its slow capacity to update makes it look persistent and unaffected immediately by changes in the environment. This capacity is emphasized in models that decrease their learning rate over time or models that have a constant but very small learning rate, which has been shown to better cope with noisy uncertain environments. Hence, in case of devaluation or contingency degradation, an MF system mainly replicates what it has been extensively trained to do. Its limited requirements in term of computational capacities also make it consistent with observations that limited attention or reduced memories capacities favour Habitual behaviours [Ott+13; Foe+06; Ger+14].

The exact interaction between the two systems is however still an open question [SD12; Dol+12]. Various propositions have been made, which we classified in two categories: winner-take-all models and integration models.

Note that alternatives combining RL systems with other systems (e.g. with a working-memory system [CF12] or with an action sequence learner [DB12a; DB13]) or based on principles at the biological-level exist [Mar+11; Arl+04; AG00; Che+13; CS11; Bal+13].

For example, some have argued that the parallel between Model-Free and Habits was misleading and suggested that Habitual behaviours would better be explained by a system that learns sequences of actions [DB12a; DB13].

Winner-take-all models

We define as winner-take-all models, those where only one system (or value from one system) drives the behaviour at a time such that systems *compete* for the control of the expressed behaviour [Daw+05; Ker+11; Pez+13; Doy+02; Cha+05; KNR09; Dol+10]. This competition usually relies on some predefined criterion.

For example, Daw et al. [Daw+05] proposed a competition based on uncertainty. That is, the system in control should be selected given how accurate it is estimated to be in the current situation. At early stages, an MB algorithm is usually more accurate than an MF system. After a single successful trial, an MB algorithm has already acquired enough information to build a full policy. The more complex the problem, the more trials it takes an MF system to propagate the final reward value to the initial state, and finally provide a first useful policy. However, after extensive training in a stable environment, an MF algorithm is as accurate as an MB system. Depending on the metric used to compute uncertainty (in this case a Bayesian method [Dea+99]), it is even possible for the MF system to be considered more accurate than the MB one. Hence, at early stages the MB system guides behaviour but is replaced by the MF system on the long run, in accordance with experimental observations [AD81; Ada82].

Keramati et al. [Ker+11] proposed another criterion: the speed/accuracy trade-off. Contrary to the former model, the competition is not at the system level but at the action level. The main idea is to have Model-Free values defined as Gaussian probability distributions, such that they embed accuracy information with their variance. Hence, for an action which value is supposed to be very inaccurate, the MB system is used to infer a better value using tree search methods. Hence, the criterion defines for each action whether the Model-Free value is accurate enough or whether it is worth to invest some computational time and energy to rely on the more expensive MB system (see also [Pez+13]). At early stages, the MF system is not accurate and the global model finds it worth to use the MB system for most of the actions. However, after extensive training, the MF system accuracy is good enough to avoid slow and costly tree searches. Once again, this is consistent with experimental observations that Model-Based approaches dominate early stages of training but seem overtaken by Model-Free processes after extensive training [AD81; Ada82].

Some models [Doy+02; Cha+05; Dol+10; Cal+12] suggest to combine reward prediction errors and/or estimated values into an estimated performance criterion. For example, the model of Dollé et al. [Dol+10] uses a mechanism that memorizes which system has been the most efficient in each regions of a maze, and is updated through some kind of RPE signal. The model successfully reproduced the behaviour of rats in navigation tasks in which a cue-guided strategy and a path-planning strategy were shown to interact [Pea+98; Dev+99].

Other models propose to combine multiple identical algorithms but parametrized differently or specialized on sub-parts of the problem, hence providing a repertoire of behaviours that can be selected at each step to efficiently solve the current task [Doy+02; CK12; Kha+06; Bal02]. For example, by varying the discount rate, one can have systems specialized on short term versus long term horizons. In such cases, the selection criterion is often based on the estimated performance.

Winner-take-all approaches imply to define whether all systems are active at a time, and whether they all are revised based on the current behaviour. In some cases, one system might actually be very costly to update and block the current process (e.g. tree-searches in complex problems may become intractable without some pruning method [Daw+05; Ker+11]). In other cases, it might be hard to estimate the performance of a system on a current situation without actually running the system multiple times on it [Dol+10]. How systems can be be compared, i.e. to find a common currency, is not often straightforward [Dol+10]. Finally, such approaches usually result in each system expressing a particular and independent behaviour, even if selecting a different system at each step might lead to observe some kind of mixed behaviour.

In the present work, we focus on the inter-variability of some behaviours [Fla+11b]. However, while they seem very distinct at a higher level, they share some similarities in their properties (e.g. learning rate) or aspects (e.g. consumption-like engagement) that might be difficultly explained by winner-take-all models.

Integration models

We regroup under the term integration models, those that *integrate* information from multiple systems to guide the behaviour. For example, systems can blend multiple value functions into one that will be used to select the final decision. Another possibility is to have a dominant system that uses a second system at a certain level of its process.

Huys et al. [Huy+12] proposed a computational model where an MB system guides the behaviour using tree search methods, but some paths in the tree are pruned and replaced by values computed in a Model-Free way to limit the time and space capacity requirements. If such capacities are sparse (by stress or overload tasks), the MF dominates, i.e. it produces a Habitual behaviours. If such capacities are abundant, then one can deal with big trees and make important lookahead and adapt to changing situations quite quickly, as in Goal-Directed behaviours. This is again consistent with experimental data regarding the impact of limited working-memory capacities [Ott+13; Ger+14] or distractions [Foe+06] on the general behaviour.

In a different way, Dezfouli and Balleine [DB13] proposed that the behaviour is controlled by an MB system, where not only actions but also sequences of actions can be selected (similar to options in Hierarchical RL literature [Sut+99; Bot+09]). Such sequences would be learned by another system relying on the Model-Based values and using some kind of TD-Learning principle to compute the advantage of grouping actions into sequences (it is not an MF system). Hence, once launched in a sequence of actions, the behaviour would be Habitual. It is also usually argued that such options develop over time to fasten reaction times and is once again consistent with the shift from Goal-Directed to Habitual behaviours over time [AD81; Ada82].

Gershman et al. [Ger+14] proposed that the behaviour is controlled by an MF system, but which may be trained *offline* by an MB system, much like in the DYNA architecture [Sut90]. The model successfully explained how subjects might change their choice preferences in the first step of a 2-steps task after observing the result in the second step given their cognitive load. Participants that were asked to perform a demanding secondary task during the second step were less able of retrospection capabilities than participants that did not have to perform it. This is interpreted as if the MB system may only be able to train the MF system in resting periods [Ott+13; Ger+14].

Finally, other studies suggested that the systems work in parallel in a *flat* collaborative way. More precisely, each system learns its own action-value functions, which are afterwards integrated into a unique action-value function on which to decide what action to do next [DB02; Day+06; Glä+10; Daw+11]. This kind of model accounted for the choice preferences of humans in a stochastic 2-steps task, especially in explaining how such choices were revised after the action taken did not lead to the most probable state [Daw+11]. Such integration produces a behaviour that is neither fully Model-Based nor fully Model-Free, and is a mix between both systems. Any resulting behaviour embeds some aspects and properties of both systems. Depending on how the integration is made, one can make one system more influential than the other in the decision, which may explain some observed inter-individual variabilities of particular interest in our work [Fla+11b; Mey+12]. This is the integration approach we followed for our computational model [Les+14b; Les+14a; Les+de]. However, contrary to the currently described models, we do not only focus on instrumental conditioning but also on Pavlovian conditioning.

3.3 Pavlovian conditioning

Biologically important objects (e.g. food) automatically elicit some kind of responses in animals (e.g. salivating). They are labelled as unconditioned stimuli (US) and unconditioned responses (UR). Neutral stimuli (e.g. a light or a sound) do not usually elicit such responses, however when repeatedly paired with a US, the animal tends to develop a conditioned response (CR) towards them, CRs being most of the time similar to the UR. In such experiments, neutral stimuli are initially defined as conditioned stimuli (CS). Such CRs will remain for some time even in the absence of the contingent US. Pavlovian conditioning experiments study the development of such responses, especially their shapes, intensities and properties.

This property of stimuli to elicit responses can lead to very complex phenomena, for example when involving the concurrency of multiple stimuli, different types of contingencies, temporal properties or the involvement of memories. In this section, we discuss a subset of such phenomena (Section 3.3.1), linking them to reinforcement learning concepts when possible, or highlighting the difficulties faced when using this framework (for a deeper review of the known phenomena, we refer the reader to a dedicated special issue [AS12]). We discuss some neural correlates of Pavlovian conditioning (Section 3.3.2). Finally, we list some of the existing computational models to show their diversity and their distance with respect to reinforcement learning (Section 3.3.3). The present thesis develops a model that is a way of conciliating some properties shared by these dedicated models, to account for a number of phenomena [AS12], with the RL framework.

3.3.1 Phenomena

The simplest and oldest form of Pavlovian conditioning, *autoshaping*, was discovered by Pavlov [Pav27]. The name of the field itself is often confounded with this phenomenon. In Pavlov's experiment, the salivation of a dog (UR) was measured before the consumption of some food (US). Interestingly, the delivery of food was always preceded by a sound (CS) which after enough repetitions started to elicit salivation (CR) on its own (Figure 3.3). The phenomenon was gradual as CRs (the production of saliva) at CS time increased up to a significant level over repetitive conditioning trials. It is interpreted as if the animal learned to associate the neutral stimulus to reward delivery and started to react to it. As a good predictor of reward, the animal could anticipate the US from the CS and prepare for it. This phenomenon and most of the other phenomena subsequently described were replicated in multiple species and proved to be persistent with various conditioned stimuli and rewards, positive or negative [Dom14; BO09].

One must note that it has been observed that CRs might vary between individuals, in their intensity but also in their shapes, for example rats might either approach and engage with the CS (e.g. lever) or approach and engage with the magazine where food would be delivered [Fla+11b]. Such inter-individual variabilities are of particular interest for the present work as they are challenging to some current models of the literature that do not distinguish between multiple possible responses or which always produce one kind of response (Chapter 5).

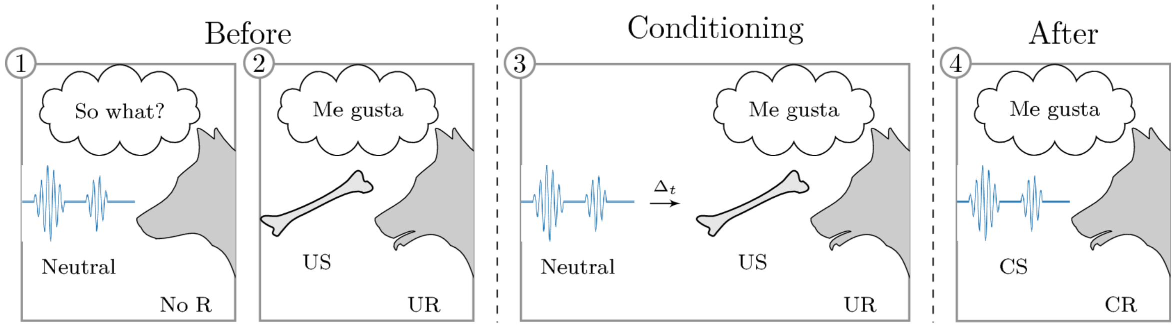


Figure 3.3: Illustration of Pavlovian autoshaping. *Before conditioning, the neutral stimulus does not produce any particular response ①. The Unconditioned Stimulus (US), e.g. some food, produces an Unconditioned Response (UR) ②, e.g. salivation. During conditioning, the neutral stimulus is paired sequentially with the US, which continues to produce a UR ③. After conditioning, the neutral stimulus elicits a Conditioned Response (CR) on its own (in extinction), thereby becoming a Conditioned Stimulus (CS) ④.*

Conditioning is usually assessed *in extinction* [Ski38; Pav27; Del04; Ger+13a], that is in a subsequent session where the CS is presented alone without subsequent presentation of the US. Conditioning is successful if animals keep producing CRs for some time in the absence of the US. The extinction procedure consists in repeating such protocol multiple times until animals completely stop to produce CRs. As its acquisition, the extinction of a CR is progressive.

From these phenomena, it is commonly accepted that the animal's behaviour results from reinforcement learning processes to build and possibly break associations between US and CS [Niv09; Mai09]. By repeatedly experiencing contingencies between CS and US, some kind of value/properties is propagated from the US to the CS or strengthens links between the two.

Stimuli competition

Autoshaping and extinction phenomena suggest that some kind of associations can be built and broken between CS and US, but do not provide information regarding a possible competition between multiple potential CSs. Phenomena such as blocking [Kam67; Hol+14], overexpectation [LN98; KM96; Res99] or overshadowing [Rey61; Mac76] spread light on this aspect of Pavlovian conditioning.

The blocking phenomenon [Kam67; Hol+14] can be observed when trying to condition a neutral stimulus S_2 (e.g. light) in the presence of an already conditioned stimulus CS_1 (e.g. sound) (Figure 3.4). After an autoshaping phase with CS_1, the second stimulus S_2 is presented in parallel to it before the presentation of the US. One could expect that S_2 would also become a CS as it is contingently paired with the US, however, when tested alone, it does not produce a CR. Hence, it is said that CS_1 blocked S_2 from becoming a CS.

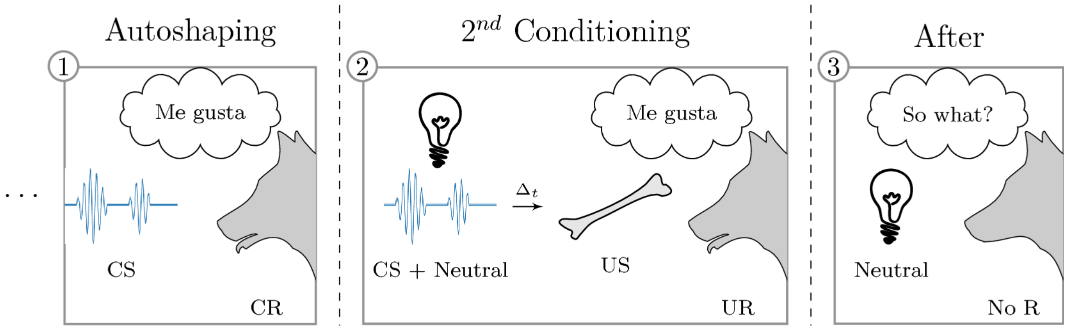


Figure 3.4: Illustration of Pavlovian blocking. *After an autoshaping phase ①, a second neutral stimulus is presented in parallel to the CS and the US, which continues to elicit a CR/UR ②. After this second conditioning phase ③, contrary to what would be expected from autoshaping, the neutral stimulus fails to produce a CR, i.e. it remains neutral.*

In an overexpectation procedure [LN98; KM96; Res99; McD+14], two stimuli are conditioned independently with the same reward. When presented alone, they elicit a certain level of CR. Interestingly, when presented together in a subsequent session where only one reward is available, they first start to elicit a higher level of CR. It is interpreted as if the animal was, to a certain extent, summing its expectations about rewards [Res99]. Furthermore, if subsequently tested again alone, the two CSs elicit a lower level of CR than originally produced, as if their compounded presentation reduced their individual level of producing CRs.

In an overshadowing [Rey61; Mac76] procedure, an autoshaping procedure is conducted where the CS is a compound of two stimuli. When presented alone, each CS produces a lower level of CR than when presented in compound. Furthermore, one of the stimuli usually elicits a stronger level of CR than the other. It is said that one stimulus overshadowed the second during conditioning.

These three phenomena reflect that a compound of stimuli should not be considered as a single stimulus. Each stimulus seems to acquire its own motivational property, which is revised in the presence of others. Hence, these kinds of phenomena do not fit well with the classical MDP framework (Section 2.2) where the whole situation – the compound presentation of the stimuli – would be defined as a single state, without taking into consideration the underlying structure and therefore without propagating value modifications to situations, i.e. other states, where stimuli are presented alone.

Propagation

Sensory preconditioning [Bro39; RR72; Pol+13] and second order conditioning [RR72; HR75; Jar+06; Mol+12] procedures consist in making an animal produce a CR in the presence of a stimulus that has never been paired directly with a reward, but instead paired with another CS that did produce a CR.

In a second-order conditioning procedure [RR72; HR75; Jar+06; Mol+12], the animal first undergoes a classical autoshaping experiment resulting in a CS_1 (e.g. light) that elicits a CR. In a second phase, this CS_1 is preceded with another CS_2 (e.g. sound). Finally, in a final test phase, one can observe that the CS_2 now elicits a CR. Hence, the animal propagated the link or value between CS_1 and US to CS_2. Note that time is important, since if presented together, the first stimulus might have blocked the second from becoming a conditioned stimulus [Pol+13; Mol+12; MM14b]. The procedure is illustrated in Figure 3.5.

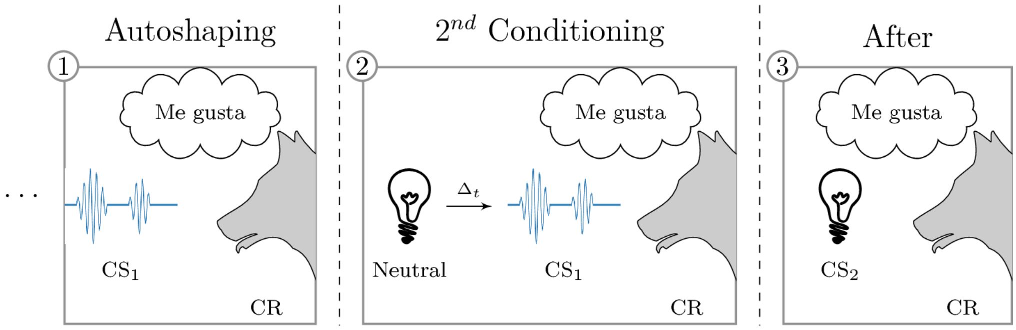


Figure 3.5: Illustration of second-order conditioning. *In a first phase, the animal undergoes an autoshaping procedure and is conditioned to a stimulus CS_1 ①. In a second phase, the CS_1 is paired with a second neutral stimulus ②. In a third test phase, the neutral stimulus is confirmed to be a conditioned stimulus CS_2 ③.*

In a sensory preconditioning procedure [Bro39; RR72; Pol+13], the two first phases of the second-order conditioning procedure are reversed, such that the link between the two initially neutral stimuli is learned first. As a result, the neutral stimulus that was never paired with the reward still elicits a CR in the final test phase.

The first procedure [RR72; HR75; Jar+06; Mol+12] suggests that any conditioned stimulus can, to some extent, work as a reinforcer to subsequently condition other stimuli. Hence, there can be a propagation of the animal “anticipation” from the most proximal to the most distal stimuli of a chain that is expected to lead to a reward, which is well accounted for by value functions in reinforcement learning processes. Indeed, the optimal path in the MDP is learned by propagating the value of an immediate reward $V(US)$ to the immediately preceding state CS_1, which value $V(CS_1)$ can in the second phase be also propagated to its preceding state CS_2. The value V over a stimulus is suggested to reflect the level of CR it produces. Hence, as soon as it has some value, a stimulus may propagate it to others.

The second procedure [Bro39; RR72; Pol+13] suggests that the link between the two stimuli was indeed learned in the first phase – while no reward was present, thus relying on some sort of latent learning –, otherwise the CS_2 would never have elicited a CR by itself. In such case, a pure MF algorithm would fail, as the absence of reward in the first phase keeps the values of CS_1 and CS_2 null. In the second phase, the value acquired by CS_1 cannot be propagated to CS_2 as this stimulus is not present. Hence, entering the test phase, the value of CS_2 should be null and, contrary to observations, no CR should occur. An MB algorithm, on the contrary would have learned the contingency between CS_1 and CS_2 in the first phase, as part of its internal model and used it to propagate the value of CS_1 to CS_2 before (or at) the test phase. Hence, MB processes are better suited to explain such data. Subsequently, the presence of a CR at the very first presentation of the second CS in the third phase of the two experiments suggests that either an offline process occurred between the second and third phase (e.g. Dyna-Q) or that values could be dynamically inferred at CS presentation (e.g. tree search methods) [Tin+09; Jon+12].

Contextual informations

Blocking, overshadowing and overexpectation phenomena suggest some kind of individual processing of stimuli. However, this does not necessarily exclude that other informations are processed differently and used in the resulting behaviour. The spontaneous recovery [Res97; Res97; LW08; Ger+13a] and renewal phenomena [Ros+07; BR94; Bou04; Tho+03] provide some insights on this issue.

To observe spontaneous recovery [Res97; Res04; LW08], the animal first undergoes an autoshaping procedure and develops a CR at CS presentation. Then, the animal undergoes an extinction procedure and stops producing such CR. After some time, the animal is once again presented with the CS alone and, surprisingly, it starts to produce a CR again (Figure 3.6). Such a recovery can also be produced by presenting the US in the context of extinction (reinstatement phenomenon [RH75; BB79]) or by playing with the conditioning context (renewal).

Renewal phenomena [BR94; Bou04; Tho+03; Ros+07] occur when the context of extinction differs from the context of autoshaping and test (e.g. by a different Skinner box,

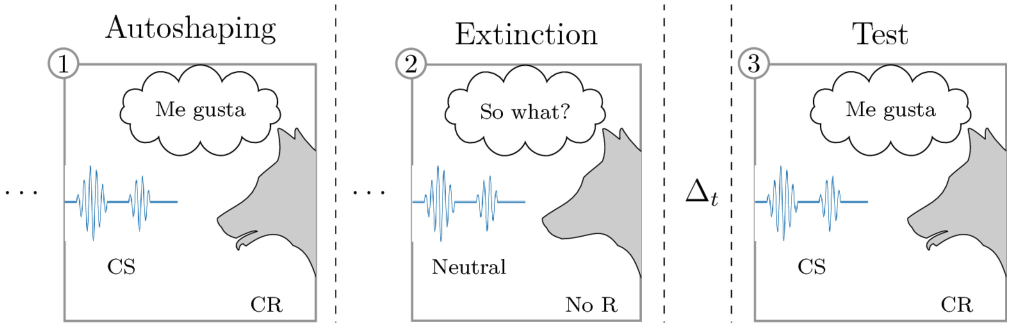


Figure 3.6: Illustration of spontaneous recovery. *The animal first undergoes an autoshaping procedure and starts eliciting a CR at CS ①. In a second phase, the animal undergoes extinction and stops producing CRs ②. After some time, the animal is once again presented with the CS, and usually spontaneously starts to produce a CR again ③.*

or different contextual cues such as coloured lights). Especially, if autoshaping is done in a context A, extinction is done in a context B and test is done in context A without a resting period (ABA renewal), the animal immediately produces a CR. It is interpreted as if the animal learned that extinction was specific to context B. If test is done in another context C (ABC renewal), the animal also produces to a lesser extent some CR, as if autoshaping was global while extinction was context-specific.

These phenomena suggest that extinction cannot be attributed to unlearning only, as spontaneous recovery shows that the CS-US association may persist despite extinction. There are various interpretations [Ger+13b; MM14a; BT14; GN12; SM07; Red+07], but all of them agree on a competition between either multiple systems, multiple contexts, or multiple memories. These phenomena also suggest that, while stimuli can compete within a same situation, there might be other mechanisms working at the context level, that is, splitting tasks into different contexts or states [Red+07; Cal+12]. The classical MDP representation, where states would embed contextual information, is well suited to explain such context-dependency. However, this representation lacks the generalization properties that would let the animal immediately produce CRs in new contexts (e.g. ABC renewal).

3.3.2 Neural correlates

While mostly put aside in the discussed phenomena, the discovery and understanding of the neural correlates of Pavlovian conditioning are an important part of the current literature (see [Niv09; Bal+08] for reviews). In this section, we describe the most influential hypothesis of dopamine as an RPE signal in the brain, and regions that were correlated with Pavlovian conditioning processes.

The Reward Prediction Error hypothesis of dopamine

Dopamine is a neurotransmitter that has been found to play a number of important roles in the human brain and is secreted in the Substantia Nigra Pars Compacta (SNc) and Ventral Tegmental Area (VTA) in strong interaction with the Basal Ganglia [TS12; Abr+14]. In a seminal paper, [Sch+97] found that the phasic dopaminergic activity in SNc and VTA evoked during Pavlovian conditioning paralleled the RPE signal involved in MF reinforcement learning algorithms. More precisely, at the beginning of conditioning, dopamine neurons fire only at US time but after sufficient training, when the CS elicits a CR, dopamine neurons fire at CS onset only. That is, there is some kind of dopamine signal that propagates from US time to CS time with conditioning. Interestingly, if the US is omitted (as in extinction), dopaminergic neurons stop firing, i.e. their activity is below baseline, at the expected US time (Figure 3.7 A). Modelling the conditioning task with the MDP framework and applying a TD-learning algorithm on it produces similar results (Figure 3.7 B).

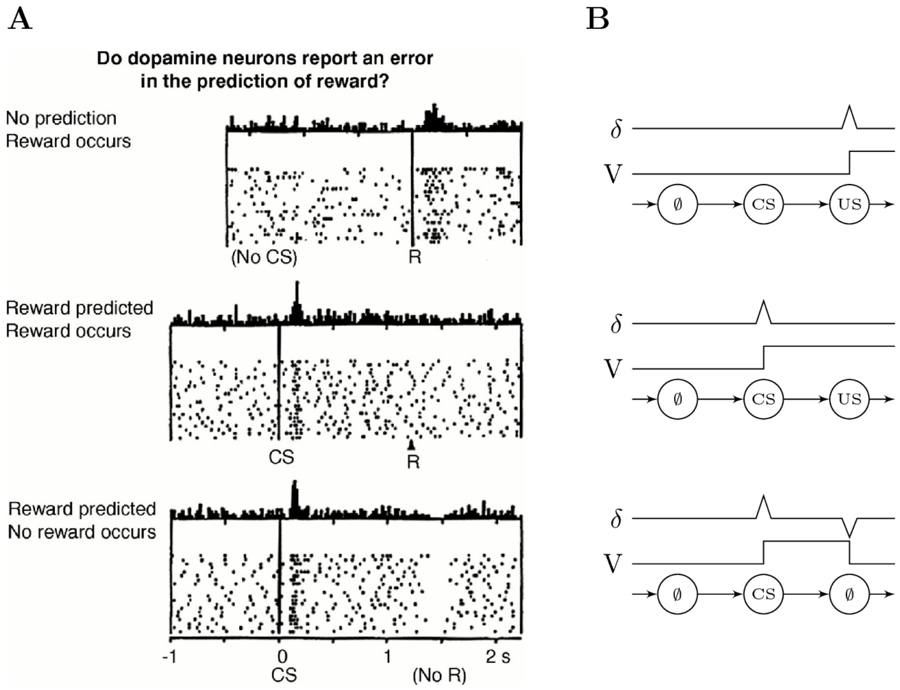


Figure 3.7: Illustration of the Reward Prediction Error hypothesis of dopamine (A) *Dopamine recordings during autoshaping and extinction. (Top) Before learning, dopamine neurons only fire at US. (Middle) After conditioning, dopamine neurons fire at CS only. (Bottom) If reward is omitted, dopamine neurons show a pause in firing at the time it was expected. Reproduction of Figure 1 of [Sch+97].* **(B)** *Reproduction of these patterns with a TD-Learning algorithm, for each phase showing the MDP used, the values currently learned V and the reward prediction error δ computed.*

Since this discovery [Sch+97], dopaminergic activity have been deeply investigated and shown to present numerous properties, a significant number of which being consistent with such hypothesis [Sch10; Sch13]. For example, the dopamine signal may encode multiple and distal rewards [Eno+11; Yam+13]. Its activity varies relatively to the quantity of rewards [BG05] and the probability of expected rewards [Fio+03]. It is sensitive to time [Fio+08] and may be context-dependent [Nak+04]. Interestingly, phasic dopaminergic activity has also been observed in instrumental tasks, matching RPEs that would take actions into account [Roe+07; Mor+06; Bel+12a; Bel+13].

Of particular importance for the present work, some studies showed that different patterns could be observed depending on the CR actually produced by animals [Fla+11b]. In an autoshaping task, where a lever was used as the conditioned cue, rats which mainly engaged with it displayed an activity similar to the one initially observed by [Sch+97], other rats which mainly engaged with the magazine during lever presentation did not have such pattern. Instead, they developed a phasic peak at both US time and CS time. Such data are challenging to the RPE hypothesis of dopamine, and unaccounted for by classical models that can only produce and explain the former pattern. The computational model developed in this work suggests an explanation for these conflicting results and a way to solve this problem with regard to the RPE hypothesis of dopamine.

Most of these results suggest that phasic dopamine is a key signal for conditioning. Since its first parallel between the RPE of TD-Learning (see Section 3.3.3), various computational models have been developed on top of an internal RPE signal that would map a maximum of the properties of phasic dopaminergic activity (e.g. [Mir+13; Roe+12; KNR09; Fio+14; Igl+13]).

One must note that such extensive analyses of dopaminergic activity have lead to an active debate regarding its exact role in the brain [Red+99a; SC02; Ber07; SC12; Eve14; Nic10; Sch10; Fio+13; Hir14; Gli11; Ste+13]. In the present work, we tend to agree with the RPE hypothesis of dopamine and the computational model developed embeds such an RPE signal.

Anatomical correlates

Similarly to instrumental conditioning, various methods (e.g. pharmacological studies, lesion studies and brain imaging) shed some light on the potential brain regions on which would rely the expression of Pavlovian conditioning. For example, functional Magnetic Resonance Imaging (fMRI) of Striatum showed that prediction error signals in the Dorsal Striatum (DS) and Ventral Striatum (VS) could be attributed to different systems [O'D+04]. Signals in Pavlovian conditioning tasks were restricted to VS, while signals in instrumental tasks were present in both VS and DS. Lesions of Orbitofrontal Cortex disrupt Pavlovian conditioning [OB07], especially [Jon+12] seem to block Model-Based capacities in Pavlovian conditioning tasks. Injection of flupentixol (an antagonist of the dopamine) blocked the expression of some CRs in rats undergoing an autoshaping experiment [Fla+11b]. Lesions of Ventral Striatum, or infusion of dopaminergic agonist within it have been shown to disrupt the learning and expression of Pavlovian approaches [Par+99; Par+02].

From these observations, it is commonly assumed that Pavlovian and instrumental conditioning rely on separate mechanisms that are mainly distributed within the Basal Ganglia, but also in the rest of the brain. Hence, it is common that models of Pavlovian conditioning do not embed instrumental aspects, and vice versa.

3.3.3 Theories and models

It is interesting to note that while reinforcement learning developed significantly through the study of Pavlovian conditioning, the formal framework of RL (which is about learning sequences of actions) is actually better suited for instrumental conditioning.

The current section tries to show the wide diversity of approaches in modelling Pavlovian conditioning through some examples of dedicated computational models. Some rely on simple RL principles, some on multiple specialized modules and some on neural network architectures. However, they share in common that their output reflects intensity of responses rather than possible sequences of actions, and process stimuli individually. The computational model developed in the present work (Chapter 5) takes advantage of the latter idea.

TD-Learning

Introduced as an extension to earlier models [RW72; SB81], the TD (Temporal Difference) model [SB87] is the straightforward application of the Temporal Difference Learning principle (Section 2.3.1). There is no action and it is assumed that the values acquired by stimuli reflect the intensity of CRs that would be expressed towards them. In its original version [SB81], values are defined over stimuli rather than states, and the value of a situation (necessary to compute the RPE in a traditional way (Equation 2.11)) is computed as a sum of the values of stimuli that compose it, that is $V(s) = \sum_{cs_i \in s} V(cs_i)$.

This TD model can replicate some Pavlovian conditioning phenomena, such as blocking effects or secondary conditioning [Sut90; BM98; Niv09; Lud+12]. Depending on the MDP representation used (e.g. by making intermediate states unique), it might be able to represent delayed conditioning (when the CS offset and the US onset are separated by some time) [BM98; Lud+12; KNR09]. Purely based on learning, it fails to show any recovery effect or pre-exposure effects. It also fails to account for phenomena that suggest Model-Based aspects [DB14; Jon+12]. This model is actually well suited to explain the core aspects of known phenomena but shows its limits as soon as subtleties are introduced [Lud+12]. However, it is interesting on the biological side, as phasic dopaminergic activity has been observed to match the RPE signals of this model in a significant number of cases [Sch+97; Sch07; Sch10] (Figure 3.7).

SLG model

Schmajuk et al. [Sch+96] have been developing the SLG (for Schmajuk-Lam-Gray) model over several years [Sch+96; SL06; LS08], extending it to account for a significant number of phenomena, especially variants of the main Pavlovian phenomena (most of which will not be discussed in this chapter) that imply for example delayed responding, inhibition or temporal precision in CS presentation.

The SLG model is composed of multiple modules (Novelty, Attention, Feedback, Inhibition and Model) that interact with each other to return the intensity of the CR and OR (Orienting Response) that should be expected from the animal, given the presentation of one or multiple CSs. The feedback system provides the necessary tools to handle delayed conditioning, by maintaining a trace of oldest events. The attentional system shifts the animal's attention towards either salient stimuli or novel stimuli. The novelty system computes the global novelty of the current situation which is propagated to other modules of the model. The inhibition system accounts for some observations that novelty might inhibit conditioned responses at first, but not orienting responses. Finally, the last system is the core of the model, where links are built between CSs and US providing some kind of valued map of the world (Figure 3.8).

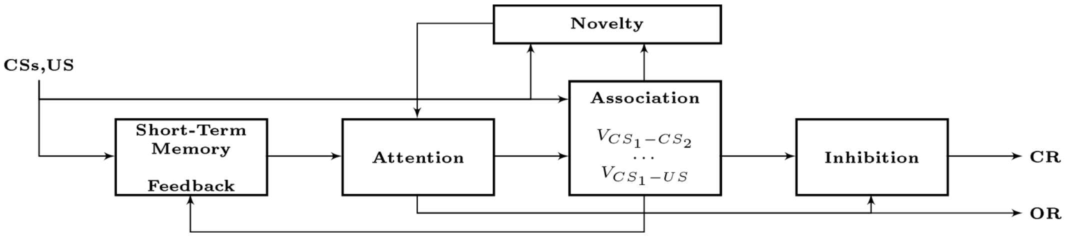


Figure 3.8: Illustration of the SLG model. *The model possibly takes multiple CSs and a unique US as input over time and through a set of specialized modules is able to express as output the intensity of the associated CR and of the associated oriented response (OR). Inspired from Figure 1 of [Sch+96] and Figure 1 [SL06].*

This model and its extensions can account for a significant number of phenomena and their derivatives (see Table 1 of [KS12]).

The comparator hypothesis

The comparator hypothesis and its extensions [MM88; Den+01; SM07; MM14a] are built on the idea that the sole contiguity of a CS and a US, or two CSs build direct associations between them. The complex phenomena of Pavlovian conditioning occurs only at the expression time of CRs, due to some competition process (see Figure 3.9). At CS onset, memories of the *comparator stimuli* (other CSs associated to CS) are retrieved and compete in the control of behaviour. While the CS directly activates a (direct) representation of US, it is compared to (indirect) representations of the US activated by the *comparator stimuli*. For example in blocking, when presented alone, S_2 (the light) retrieves a memory of CS_1 (the sound) that activates a stronger indirect representation of US than the direct representation, and results in the absence of CR. If CS_1 is presented alone, the indirect representation of US activated by S_2 is weaker and a CR is produced. Each association can be modulated by second-order *comparator stimuli*, which allows the model to explain complex stimuli-competition phenomena.

This model and its extensions can explain complex stimuli-competition phenomena but is limited on some other aspects such as extinction and renewal (see Figure 15 of [SM07]).

Latent Cause Theory

In contrast to classical associative theories where animals are hypothesised to learn CS-US associations (e.g. the comparator hypothesis), Courville et al. [Cou+04] hypothesised that

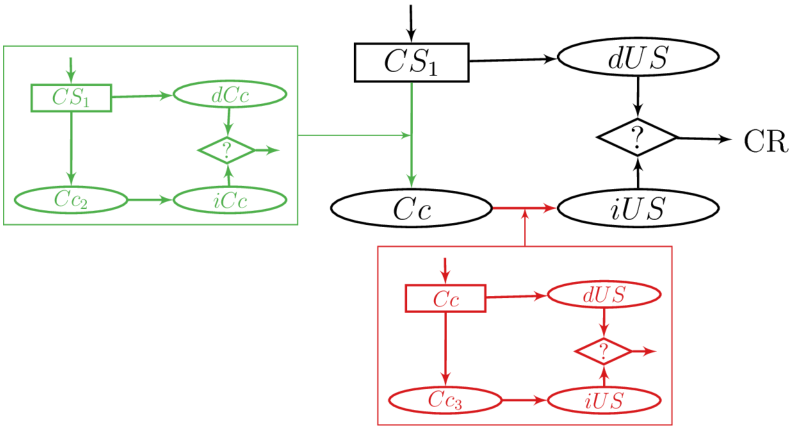


Figure 3.9: Illustration of the extended comparator hypothesis. *Presented with a CS_1, it elicits a direct representation of the US (dUS), and an indirect representation of the US (iUS) through a comparator stimulus Cc. The strength of each indirect link can undergo the same process, such that for example a CS_1 can elicit a direct representation (dCc) and an indirect representation (iCc) of a comparator stimulus, and so on. Inspired from Figure 1 of [MM14a].*

animals instead attempt to infer a generative model of the world. In the Latent Cause Theory [Cou+04] or related alternatives [GN12], it is assumed that any observation arises from an invisible cause and co-occurring events share this same hidden cause. Links are therefore created between a probable cause and an observation, hence stimuli are only connected through a latent cause. For example, the CS (e.g. sound) and the subsequent US (e.g. food) of a classical Pavlovian task are to be seen as co-occurring because of some hidden cause (e.g. the procedure defined by the experimentalist). More precisely, what is learned is actually probabilities $P(\text{observation}|\text{cause})$ (see Figure 3.10 for a graphical presentation of the theory).

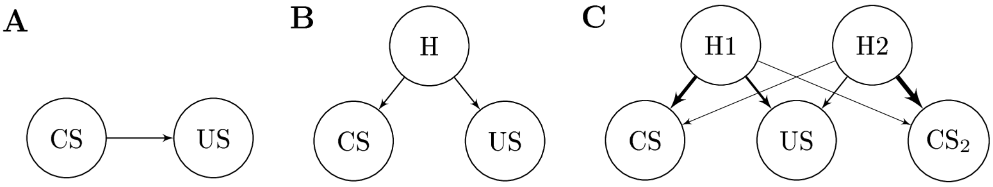


Figure 3.10: Visual interpretation of the Latent Cause Theory. **(A)** *Classical associative theory: CS and US are associated.* **(B)** *Latent cause theory: CS and US are associated to their shared latent cause. The first layer is composed of hidden variables (latent causes). The second layer is composed of observations.* **(C)** *Illustration of the probabilistic properties of the Latent Cause Theory.*

Using Bayes theorem

$$P(cause|obs) = \frac{P(obs|cause) \times P(cause)}{P(obs)} \tag{3.1}$$

one can first compute the conditional distribution over the possible (new) causes given the current observations $P(\text{cause}|\text{obs})$, and then infer new observations to be expected. Furthermore, new observations lead to the incremental revision and improvement of the model.

Depending on the exact implementation of the Bayesian processes involved, especially regarding how latent causes are selected or newly created (e.g. through a discriminative versus a generative process), one might be able to account for different sets of Pavlovian phenomena (e.g. stimuli competition [Cou+06], contextual properties [GN12]).

3.4 Pavlovian-instrumental interactions

Pavlovian and instrumental conditioning are usually studied separately, assuming that protocols are sufficient to solicit one conditioning and not the other. But, while Pavlovian and instrumental conditioning have been shown to rely, at least partially, on different brain mechanisms, their complete separation is not so clear [Yin+08; Mee+12; LO12; Mee+10; Mai09].

Some Pavlovian phenomena have counterparts in the instrumental world, for example, recovery phenomena [Nak+00; Tod+12; Bou+12], overexpectation [LN98] or contextual effects [Bou+14; Mar+13]. Some Pavlovian procedures use conditioned stimuli (e.g. lever) that are used as operant objects in instrumental task [Fla+11b]. It is often the case that instrumental tasks embed cues commonly used as conditioned stimuli (e.g. sounds) to inform animals about the different phases of the task. Finally, some phenomena clearly emphasize that they can easily interact in a very tight and complex way.

These phenomena have recently been the focus of an increasing number of studies [Lov83; Hal+01; HG03; CB05; CB11; Tal+08; Car+13; Hol+10]. In this section, we briefly present some of the major interaction phenomena (Section 3.4.1), their neural correlates (Section 3.4.2), and the few computational models that have been developed to account for them (Section 3.4.3).

3.4.1 Phenomena

Phenomena that have been suggested to arise from Pavlovian and instrumental interactions can emerge in different ways: by preceding a classical instrumental protocol with a Pavlovian one, or vice versa, by combining instrumental and Pavlovian protocols or by training subjects separately on Pavlovian and instrumental protocols and subsequently testing the result in a combined protocol. Here we list phenomena that have been effectively listed as Pavlovian-instrumental interactions. New interaction phenomena may still be discovered yet, or older one reinterpreted as interactions.

Conditioned Reinforcement Effect (CRE)

It has been shown that an initially neutral stimulus that has been conditioned in a Pavlovian process can subsequently be used as the desired outcome of an instrumental task, such that animals will actively engage in a task to make them appear and interact with them [Wil94b; Ski38; RF09].

For example, in a study of Robinson and Flagel [RF09], rats first underwent a classical autoshaping procedure with a lever as CS. In a subsequent phase, rats were presented with an active and an inactive nose port. Nose poking into the active port resulted in presentation of the lever for 2 seconds without subsequent reward delivery, whereas poking into the inactive one had no consequence. The authors observed that rats significantly preferred the active nose port to an inactive one, clearly suggesting that the CS became actively desired by the animal during the Pavlovian phase.

This shows that the properties/values acquired by stimuli during Pavlovian conditioning can be subsequently used in an instrumental conditioning process and definitely impact it.

Pavlovian-Instrumental Transfer (PIT)

A major example of a Pavlovian and instrumental conditioning interaction is the invigoration that a Pavlovian CS can have on a instrumental action, an effect named Pavlovian-Instrumental Transfer (PIT) [CB05; Hol04; Tal+08; Cor+07; CB11; Bal94; CJ07; Hol+10; Huy+14; Hal+01]. This phenomenon has been the focus of more and more studies in the past few years, as it seems to deeply contribute to addictive behaviours, where cues seem to take control over rational behaviours. Many species, including humans show PIT effects [Hol+10; Huy+14; Tal+08].

PIT can classically be observed in rats with the following 3 steps procedure. Rats are first trained to associate some sound with the delivery of some food (Pavlovian procedure). Subsequently, rats are trained to press a lever to be rewarded (instrumental procedure) on some reinforcement schedule (i.e. multiple presses or some delay is needed for reward to be delivered). Finally, once again presented with the lever (in extinction), rats show a higher number of presses on the lever when the sound is concurrently played relative to when it is not [CB05; Cor+07; Hal+01; Hol04].

PIT is actually not a unitary effect and must be further divided between Specific PIT and General PIT, depending on whether the rewards used in the Pavlovian and instrumental procedures are similar or different. Especially, lesions of the core and shell of the nucleus accumbens can block Specific PIT and General PIT respectively [CB05; CB11].

The PIT effect suggests that what was passively learned during a Pavlovian procedure directly impacts actions in a subsequent instrumental task. Note that some experiments combining overexpectation and instrumental tasks [LN98] also suggest that the PIT effect could be affected by other Pavlovian phenomena. Furthermore, while General PIT suggests a process where the sole presence of a reward, whatever its identity, is sufficient, Specific PIT suggests a process that takes the identity of the reward into consideration. Hence, it is often seen as if General PIT could depend on some MF system, where the computational principles do not allow the system to keep information regarding the identity and only focus on values. On the other side, Specific PIT could depend on some MB system, where the internal model makes it possible to keep track of the identity of the currently investigated reward [Jon+12; Cla+12].

Negative auto-maintenance

In a *negative automaintenance procedure*, animals are presented with stimuli that precede food delivery (exactly as in an autoshaping procedure). However, if they develop the CR usually observed under autoshaping, reward is omitted. In the case of pigeons [WW69; San+06], they are presented with a colored key light and any peck on that key light terminates the trial without rewards. The experimental setup is illustrated in Figure 3.11.

It has been observed in multiple species [DW77; GR73; Kil03; Woo+74; Loc+76; Loc+78; O'C79; GH72; Kon48; She65] that animals are unable to completely block CRs, such that they lose a significant amount of reward during the process.

Of particular interest to the present work, results vary on how this procedure is ineffective at blocking a CR expression. More precisely, it has been observed in pigeons that some are more efficient at refraining from pecking than others [WW69; San+06]. It suggests once again some form of inter-variability in the population (see Chapter 7).

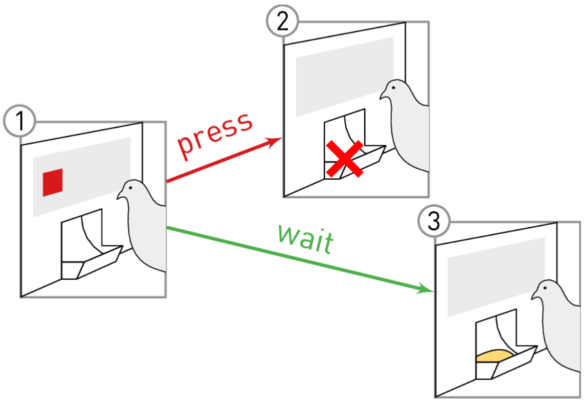


Figure 3.11: Negative-automaintenance procedure. *The pigeon is put in a Skinner box where a key light is subsequently turned on ①. If pigeon pecks at the key light, it is immediately turned off and no reward is delivered ②. If it waits for a small period of time (8s) after which the key light is turned off, reward is delivered in the food cup ③. Trials are separated by some inter-trial interval.*

According to multiple studies [Day+06; Loc+78; San+06], this phenomenon confronts Pavlovian processes and instrumental ones. Conditioned responses develop because of the contingency between the conditioned stimulus and the reward (Pavlovian conditioning). We would expect pigeons not to peck as it prevents them from being rewarded (instrumental conditioning). Contrary to CRE where the Pavlovian phase seems to only contribute initially to the subsequent instrumental conditioning phase, or to PIT where the two types of conditioning seem to collaborate, this phenomenon seems to directly confront Pavlovian and instrumental conditioning.

3.4.2 Neural correlates

Current studies suggest that Pavlovian and instrumental conditioning phenomena rely, at least partially, on different brain regions, and each type of conditioning might also rely on multiple systems (Figure 3.12). However, how these systems interact remains unclear. They might still overlap on non-investigated aspects, they could be organized hierarchically, or offer multiple and redundant pathways.

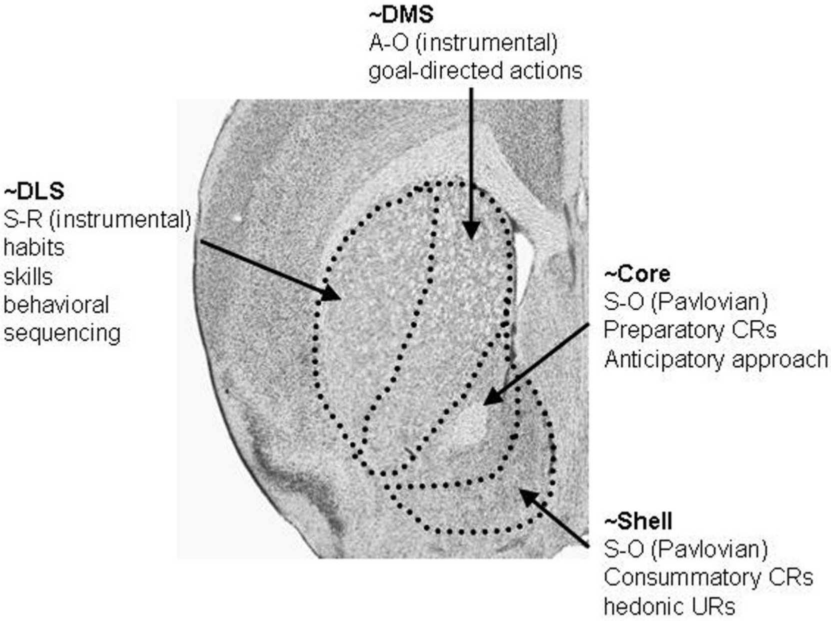


Figure 3.12: Hypothesized functional domains of the Striatum. *While illustrated by clear separations, these regions are anatomically continuous and approximation of what are commonly known as Nucleus Accumbens Shell and Core (Ventral Striatum), Dorsomedial Striatum (DMS) and Dorsolateral Striatum (DLS). Reproduction of Figure 1 of [Yin+08].*

3.4.3 Theories and models

With numerous models exclusively accounting for Pavlovian conditioning or instrumental conditioning, one would expect that explaining some of their interactions should be easily achieved by combining two of these models. However, these models usually rely on very distinct paradigms, making their combination neither straightforward nor natural [Mee+12].

Pavlovian models tend to describe the varying intensity of a unique response and its propagation to conditioned stimuli is usually defined as the acquisition of some kind of value [SB81; SB87; SB90; Sch+96; SL06; LS08; MM88; Den+01; SM07; Cou+04; GN12; Jam+12; Has+10]. The notion of action/reflex is almost always absent or hidden from such models. Instrumental models [Daw+05; Ker+11; DD13; DB12a; Huy+12; Bei+11; Dol+12; Daw+11; SD12], relying mainly on the RL-paradigm, focus on the acquisition of sequences of actions, where action selection is central to the process. The shapes of actions are almost always neglected from such models, and sometimes also their intensity. As situations are defined as states, the possibility to use relevant informations about combinations of stimuli is often lost.

As a result, few computational models have actually been developed to account for Pavlovian-instrumental interactions, most of them being very specific to the tasks studied [Kil03; Day+06; Huy+11; Car+13]. A difficulty of such approaches lies in the lack of knowledge regarding the integration of Pavlovian and instrumental systems [Mee+12]. In this thesis, we develop another computational model that account for the experimental data of multiple studies, some of which involving interactions.

The computational model of Cartoni et al. [Car+13] is an attempt to model and explain the separation between General and Specific PIT. It is also an interesting (and promising) attempt to extend the Latent Cause Theory [Cou+04] used for Pavlovian conditioning, to instrumental conditioning by introducing the notion of action. It suggests to add a third layer for actions in the DBN network of the Latent Cause Theory, but which would only be connected to primary rewards. It explains Specific PIT as resulting from the increase in the probability of being rewarded when the CS played is associated to the same food that is obtained by the currently available action (Figure 3.13 A). It also explains the lack of PIT effect, when the CS played is associated to a food that is reachable by a different action, as a necessary exclusion of latent causes due to the necessity of actions (Figure 3.13 B). Finally, it explains General PIT as some sum between multiple expected rewards (Figure 3.13 C). However, the model is yet at an early stage of investigation and would difficultly generalize to other instrumental experiments.

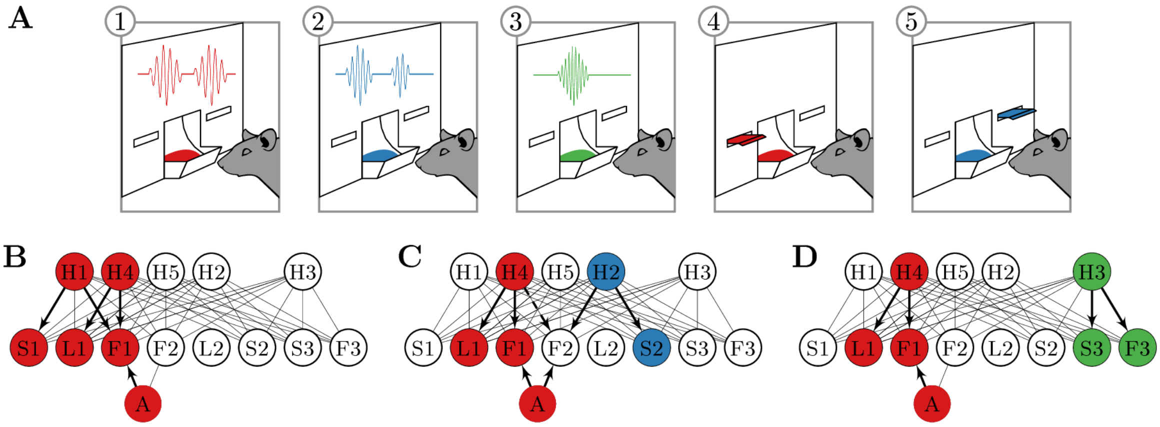


Figure 3.13: Illustration of the model of Cartoni et al. [Car+13]. (A) *Illustration of the conditioning phases. The Pavlovian conditioning phase pairs 3 different CSs with 3 different foods ①②③. The instrumental conditioning phase pairs 2 levers with 2 of the food used in the Pavlovian phase ④⑤.* **(B)** *Specific PIT appears when the CS and the lever lead to the same food ①④.* **(C)** *Absence of PIT is observed when the CS and the lever lead to different food, the CS leading to a food that would have been reachable by another action ②④.* **(D)** *General PIT appears when no action would have lead to the different food associated with the CS ③④.*

Dayan et al. [Day+06] proposed a general computational model of interactions between Pavlovian and instrumental conditioning and took negative automaintenance as an illustration [WW69], but did not attempt to precisely fit the model to the experimental data. Interestingly, this model suggests a dual-learning mechanism where a simple

RL system (accounting for the instrumental aspect) computes action values that are subsequently biased by the second Pavlovian system before the action selection phase. More precisely, the RL system learns an advantage function (A), which computes the advantage of an action relative to the other by computing its (negative) advantage value $A(s,a) = Q(s,a) - \max_a' Q(s,a')$, that is how worse it is relative to the estimated best action. This advantage is subsequently combined with a Pavlovian impetus towards a priori defined Pavlovian actions that are assumed to be hard coded in the brain (here pecking). This impetus is equal to the averaged value of the current state $V(s)$, i.e. the more rewarding is the current situation, the more biased the behaviour should be. The combination of the impetus and the advantage value is done as the following

$$P(s,a) = (1-\omega) \times A(s,a) + \omega \begin{cases} V(s) & \text{if } a \text{ is Pavlovian} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

where $0 \leq \omega \leq 1$ is a weighting parameter that enables to vary the Pavlovian influence in the overall behaviour. The computational model presented in this work [Les+14b] is originally inspired by this model.

Interestingly, Huys et al. [Huy+11] used a computational model that combined similar mechanisms to explain the results of a specific go/no-go task combined with PIT. In their instrumental phase, the subjects had to either click on an image (approach trials), click on the opposite side of the image (withdrawal trials) or do nothing (no go). The instrumental phase was conducted in a neutral context, i.e. no background image or sound. In the Pavlovian phase, the subjects passively heard sounds or saw images and were subsequently informed of a loss or a win on their final outcome. Then, they were tested on the instrumental task but with sounds played or background images displayed behind the instrumental images. They successfully explained the inhibition impact of appetitive Pavlovian stimuli on withdrawal trials, i.e. making subjects press on the image or wait rather than click on the opposite side, and the reversed impact of aversive stimuli on approach trials.

Chapter 4

Synthesis and working hypothesis

Pavlovian and instrumental conditioning are mainly studied separately, even if some phenomena clearly suggest that they are strongly entangled in the brain [Lov83; Hal+01; HG03; CB05; CB11; Tal+08; Car+13; Hol+10; Day+06]. They are mostly accounted for by different computational models that are often incompatible and rely on different principles. Few models have been developed to account for Pavlovian-instrumental interactions phenomena [Kil03; Day+06; Huy+11; Car+13]. Furthermore, such models are specific to the task modelled or not yet validated on experimental data. Hence, the research of a unifying framework, or at least a collection of compatible frameworks, is of particular interest in the study of their interactions.

The reinforcement learning framework seems particularly interesting for the given task. It explains some basic Pavlovian phenomena at the behavioural level [SB81; SB87; Sut88; RR72; HR75; Jar+06; Mol+12] and physiological level [Sch+97; Sch98; Sch10; Sch13; Gli11; Har+14; BG05]. Furthermore, the dichotomy between Model-Free and Model-Based algorithms is well suited to explain the Habitual and Goal-Directed aspects of instrumental behaviours [Daw+05; Ker+11; DD13; DB12a; Huy+12; Bei+11; Dol+12; Daw+11; SD12]. However, in its standard form this framework fails to account for some important Pavlovian phenomena, among which phenomena involving the competition between multiple stimuli [Kam67; Hol+14; LN98; KM96; Res99; Rey61; Mac76]. Such phenomena are usually explained by other models with dedicated architectures that deal with stimuli independently [Sch+96; LS08; MM88; Den+01; SM07; Cou+04; GN12]. Hence, it is of interest to see if possible extensions of the framework would overcome some of the reinforcement learning framework limitations. In particular, using factored representations [Bou+95], classical RL systems could be revised to use information about individual stimuli. This could provide useful generalization properties that are currently lacking to the general framework. It could also allow some competition or collaboration between stimuli in the expression of the behaviour.

Some experimental data about maladaptive behaviours [WW69; San+06; GM+12] are suggested to result from the interactions between Pavlovian and instrumental systems [Day+06; GM+12]. The underlying intuition is that such maladaptive behaviours can be explained by the behaviour being biased by a Pavlovian system towards selecting suboptimal actions [Day+06; GM+12]. Hence, this would suggest that the involved Pavlovian system is suboptimal with respect to the MDP framework. Some experimental data about inter-individual differences in conditioning tasks [Fla+07; Fla+09; Fla+11b; Fla+11a; RF09; Mey+12; SR12] also suggest the coexistence of multiple systems. In these data, some individuals are more prone to focus on specific reward-related cues than others, which suggests that one system should take into consideration individual cues. The inter-individual variability might then result from the importance accorded to that system by individuals.

Taking inspiration from these yet unaccounted for experimental data, our intuition is that combining two reinforcement learning systems and extending at least one of them to factored representations, which could lead to competition between cues the resolution of which being possible through both optimal and non-optimal solutions, might actually explain some of these puzzling results. Hence, we propose to investigate whether a computational model based on these concepts would successfully replicate such results.

Chapter 5

This chapter presents the computational model developed over this PhD thesis and its first application to experimental data. This work is presented under the form of a published journal paper:

> *Florian Lesaint, Olivier Sigaud, Shelly B Flagel, Terry E Robinson, and Mehdi Khamassi. "Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations". In:* PLoS Comput Biol *10.2 (2014), e1003466*
>
> `http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003466`

It aims at revising the study of Pavlovian conditioning with the RL framework by combining it with factored representations in a dual-learning system model. It proves the interest of such a method by confronting it to experimental data unaccounted for by classical theories, especially regarding recordings that do not fit with the classical RPE hypothesis of dopamine.

It shows that a computational model composed of a Model-Based system and a Model-Free system revised to use factored representations enables to reproduce inter-individual behavioural, physiological and pharmacological differences observed in rats called sign-trackers and goal-trackers in a Pavlovian autoshaping task [Fla+11b].

Simulations suggest that the behaviour of both types of animals results from a difference in the balance of the contributions of the MB and MF systems (values integrated through a weighted sum). Sign-trackers would mainly rely on the MF system, while goal-trackers would mainly rely on the MB system. The model also explains why the acquisition of goal-tracking is dopamine-independent unlike the acquisition and expression of sign-tracking.

Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations

Florian Lesaint[1,2,*], Olivier Sigaud[1,2], Shelly B. Flagel[3–5], Terry E. Robinson[5], Mehdi Khamassi[1,2]

1 Institut des Systèmes Intelligents et de Robotique, UMR 7222, UPMC Univ Paris 06, Paris, France

2 Institut des Systèmes Intelligents et de Robotique, UMR 7222, CNRS, Paris, France

3 Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, United States of America

4 Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, Michigan, United States of America

5 Department of Psychology, University of Michigan, Ann Arbor, Michigan, United States of America

∗ E-mail: Corresponding lesaint@isir.upmc.fr

Abstract

Reinforcement Learning has greatly influenced models of conditioning, providing powerful explanations of acquired behaviour and underlying physiological observations. However, in recent autoshaping experiments in rats, variation in the form of Pavlovian conditioned responses (CRs) and associated dopamine activity, have questioned the classical hypothesis that phasic dopamine activity corresponds to a reward prediction error-like signal arising from a classical Model-Free system, necessary for Pavlovian conditioning. Over the course of Pavlovian conditioning using food as the unconditioned stimulus (US), some rats (sign-trackers) come to approach and engage the conditioned stimulus (CS) itself – a lever – more and more avidly, whereas other rats (goal-trackers) learn to approach the location of food delivery upon CS presentation. Importantly, although both sign-trackers and goal-trackers learn the CS-US association equally well, only in sign-trackers does phasic dopamine activity show classical reward prediction error-like bursts. Furthermore, neither the acquisition nor the expression of a goal-tracking CR is dopamine-dependent. Here we present a computational model that can account for such individual variations. We show that a combination of a Model-Based system and a revised Model-Free system can account for the development of distinct CRs in rats. Moreover, we show that revising a classical Model-Free system to individually process stimuli by using factored representations can explain why classical dopaminergic patterns may be observed for some rats and not for others depending on the CR they develop. In addition, the model can account for other behavioural and pharmacological results obtained using the same, or similar, autoshaping procedures. Finally, the model makes it possible to draw a set of experimental predictions that may be verified in a modified experimental protocol. We suggest that further investigation of factored representations in computational neuroscience studies

may be useful.

Author Summary

Acquisition of responses towards full predictors of rewards, namely Pavlovian conditioning, has long been explained using the reinforcement learning theory. This theory formalizes learning processes that, by attributing values to situations and actions, makes it possible to direct behaviours towards rewarding objectives. Interestingly, the implied mechanisms rely on a reinforcement signal that parallels the activity of dopamine neurons in such experiments. However, recent studies challenged the classical view of explaining Pavlovian conditioning with a single process. When presented with a lever whose retraction preceded the delivery of food, some rats started to chew and bite the food magazine whereas others chew and bite the lever, even if no interactions were necessary to get the food. These differences were also visible in brain activity and when tested with drugs, suggesting the coexistence of multiple systems. We present a computational model that extends the classical theory to account for these data. Interestingly, we can draw predictions from this model that may be experimentally verified. Inspired by mechanisms used to model instrumental behaviours, where actions are required to get rewards, and advanced Pavlovian behaviours (such as overexpectation, negative patterning), it offers an entry point to start modelling the strong interactions observed between them.

Introduction

Standard Reinforcement Learning (RL) [SB98] is a widely used normative framework for modelling conditioning experiments [SB87; Bar95]. Different RL systems, mainly Model-Based and Model-Free systems, have often been combined to better account for a variety of observations suggesting that multiple valuation processes coexist in the brain [Cla+12; SD12; Car+02]. Model-Based systems employ an explicit model of consequences of actions, making it possible to evaluate situations by forward inference. Such systems best explain goal-directed behaviours and rapid adaptation to novel or changing environments [Yin+05; SB12; Daw+11]. In contrast, Model-Free systems do not rely on internal models and directly associate values to actions or states by experience such that higher valued situations are favoured. Such systems best explain habits and persistent behaviours [Gra08; Yin+04; Daw+11]. Of significant interest, learning in Model-Free systems relies on a computed reinforcement signal, the reward prediction error (RPE). This signal parallels the observed shift of dopamine neurons' response from the time of an initially unexpected reward – an outcome that is better or worse than expected – to the time of the conditioned stimulus that precedes it, which, in Pavlovian conditioning experiments, is fully predictive of the reward [Sch98; Fio+03].

However recent work by Flagel et al. [Fla+11b], raises questions about the exclusive use of classical RL Model-Free methods to account for data in Pavlovian conditioning experiments. Using an autoshaping procedure, a lever-CS was presented for 8 seconds, followed immediately by delivery of a food pellet into an adjacent food magazine. With training, some rats (sign-trackers; STs) learned to rapidly approach and engage the lever-CS. However, others (goal-trackers; GTs) learned to approach the food magazine upon CS presentation, and made anticipatory head entries into it. Furthermore, in STs, phasic dopamine release in the nucleus accumbens, measured with fast scan cyclic voltammetry, matched RPE signalling, and dopamine was necessary for the acquisition of a sign-tracking CR. In contrast, despite the fact that GTs acquired a Pavlovian conditioned approach response, this was not accompanied with the expected RPE-like dopamine signal, nor was the acquisition of a goal-tracking CR blocked by administration of a dopamine antagonist (see also [DE10]).

Classical dual systems models [Day+06; Daw+05; Ker+11; Glä+10] should be able to account for these behavioural and pharmacological data, but the physiological data are not consistent with the classical view of RPE-like dopamine bursts. Based on the observation that STs and GTs focus on different stimuli in the environment, we suggest that the differences observed in dopamine recordings may be due to an independent valuation of each stimulus. In classical RL, valuation is usually done at the *state* level. Stimuli, embedded into *states* – snapshots of specific configurations in time –, are therefore hidden to systems. In this case, it would prevent dealing separately with the lever and the magazine at the same time. However, such data may still be explained by a dual systems theory, when extended to support and benefit from factored representations; that is, learning the specific value of stimuli independently from the states in which they are presented.

In this paper, we present and test a model using a large set of behavioural, physiological and pharmacological data obtained from studies on individual variation in Pavlovian conditioned approach behaviour [Fla+07; Fla+09; RF09; Fla+11b; MB09; DB12b; SR12]. It combines Model-Free and Model-Based systems that provide the specific components of the observed behaviours [Mey+12]. It explains why inactivating dopamine in the core of the nucleus accumbens or in the entire brain results in blocking specific components and not others [Fla+11b; SR12]. By weighting the contribution of each system, it also accounts for the full spectrum of observed behaviours ranging from one extreme – sign-tracking – to the other [Mey+12] – goal-tracking. Above all, by extending classical Model-Free methods with factored representations, it potentially explains why the lever-CS and the food magazine might acquire different motivational values in different individuals, even when they are trained in the same task [RF09]. It may also account for why the RPE-like dopaminergic responses are observed in STs but not GTs, and also the differential dependence on dopamine [Fla+11b].

Results

We model the task as a simple Markov Decision Process (MDP) with different paths that parallel the diverse observed behaviours ranging from sign-tracking – engaging with the lever as soon as it appears – to goal-tracking – engaging with the magazine as soon as the lever-CS appears – (see Figure 5.1).

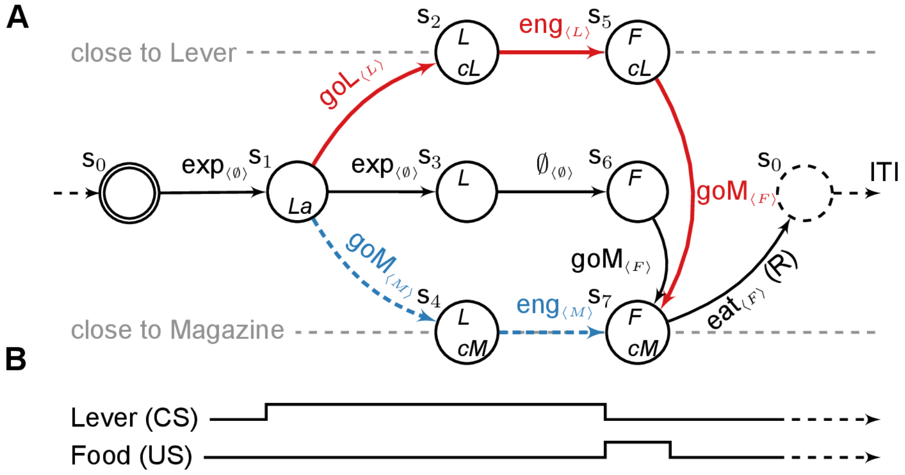


Figure 5.1: Computational representation of the autoshaping procedure. (A) *MDP accounting for the experiments described in [Fla+09; Fla+11b; RF09; Mey+12]. States are described by a set of variables: L/F - Lever/Food is available, cM/cL - close to the Magazine/Lever, La - Lever appearance. The initial state is double circled, the dashed state is terminal and ends the current episode. Actions are* engage *with the proximal stimuli,* explore*, or* go *to the* Magazine/Lever *and* eat*. For each action, the feature that is being focused on is displayed within brackets. The path that STs should favour is in red. The path that GTs should favour is in dashed blue.* **(B)** *Time line corresponding to the unfolding of the MDP.*

The computational model (see Figure 5.2) consists of two learning systems, employing distinct mechanisms to learn the same task: (1) a Model-Based system which learns the structure of the task from which it infers its values; (2) a Feature-Model-Free system where values for the relevant stimuli (lever-CS and the food magazine) are directly learned by trial and error using RPEs. The respective values of each system are then weighted by an ω parameter before being used in a classical softmax action-selection mechanism (see Methods).

An important feature of the model is that varying the systems weighting parameter ω (while sharing the other parameter values of the model across subgroups) is sufficient to qualitatively reproduce the characteristics of the different subgroups of rats observed experimentally during these studies.

To improve the matching of the following results with the main experimental data, a different set of parameter values was used for each subgroup (ST, GT and IG). The values were retrieved after fitting autoshaping data only (see Methods, Table S5.1). Simulated results on other behavioural, physiological and pharmacological data are generated with the same parameter values. While it might result in a weaker fitting of the other experimental data, this permits a straightforward comparison of results at different levels for the same simulation. Moreover, it confirms that the model can reproduce behavioural, physiological and pharmacological results with a single simulation per subgroup.

On each set of experimental data, we compare different variants of the computational

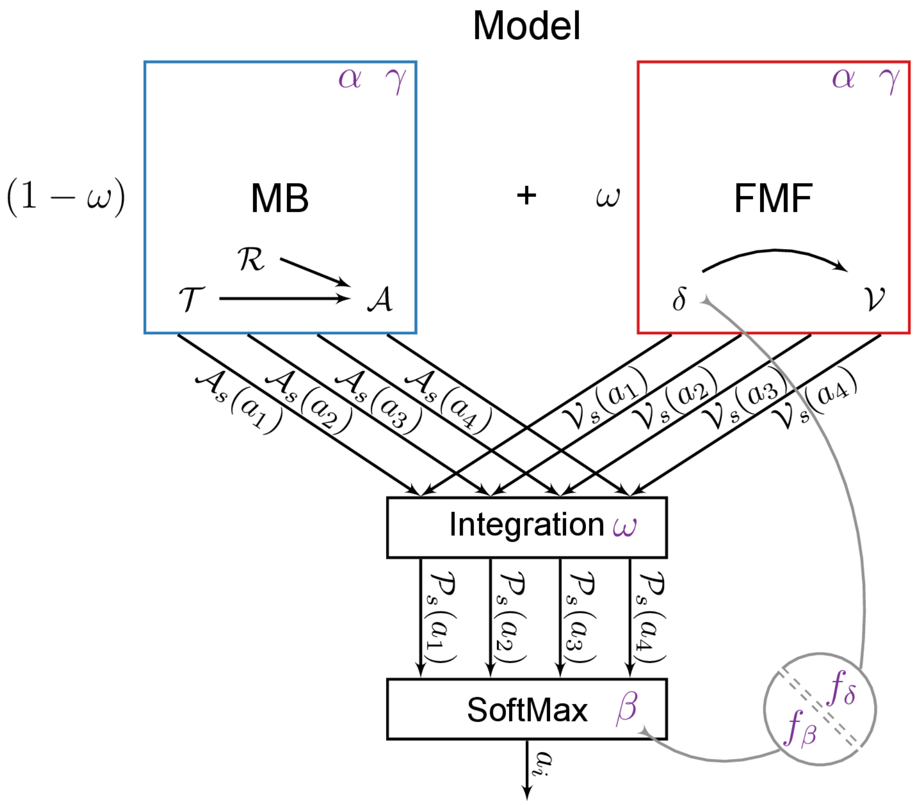


Figure 5.2: General architecture of the model and variants. *The model is composed of a Model-Based system (MB, in blue) and a Feature-Model-Free system (FMF, in red) which provide respectively an Advantage function $\mathcal{A}$ and a value function $\mathcal{V}$ values for actions a_i given a state s. These values are integrated in $\mathcal{P}$, prior to be used into an action selection mechanism. The various elements may rely on parameters (in purple). The impact of flupentixol on dopamine is represented by a parameter f that influences the action selection mechanism and/or any reward prediction error that might be computed in the model.*

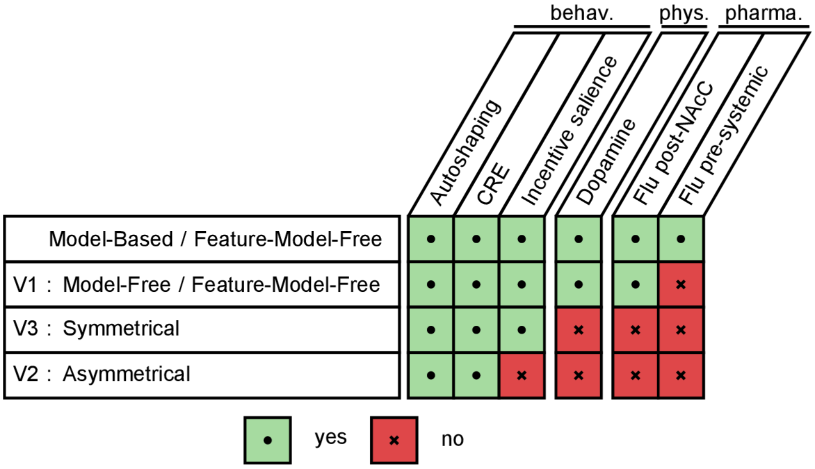
model in order to highlight the key mechanisms that are required for their reproduction. Simulation results on each data subset are summarized in Figure 5.3. The role of each specific mechanism of the model in reproducing each experimental data is detailed in Figure 5.4.

Behavioural data

Autoshaping

The central phenomenon that the model is meant to account for is the existence of individual behavioural differences in the acquisition of conditioned approach responses in rats undergoing an autoshaping procedure; that is, the development of a sign-tracking CR, a goal-tracking CR, or an intermediate response.

Based on their engagement towards the lever, Flagel et al. [Fla+09] divided rats into



	behav.			phys.	pharma.	
	Autoshaping	CRE	Incentive salience	Dopamine	Flu post-NAcC	Flu pre-systemic
Model-Based / Feature-Model-Free	●	●	●	●	●	●
V1 : Model-Free / Feature-Model-Free	●	●	●	●	●	×
V3 : Symmetrical	●	●	●	×	×	×
V2 : Asymmetrical	●	●	×	×	×	×

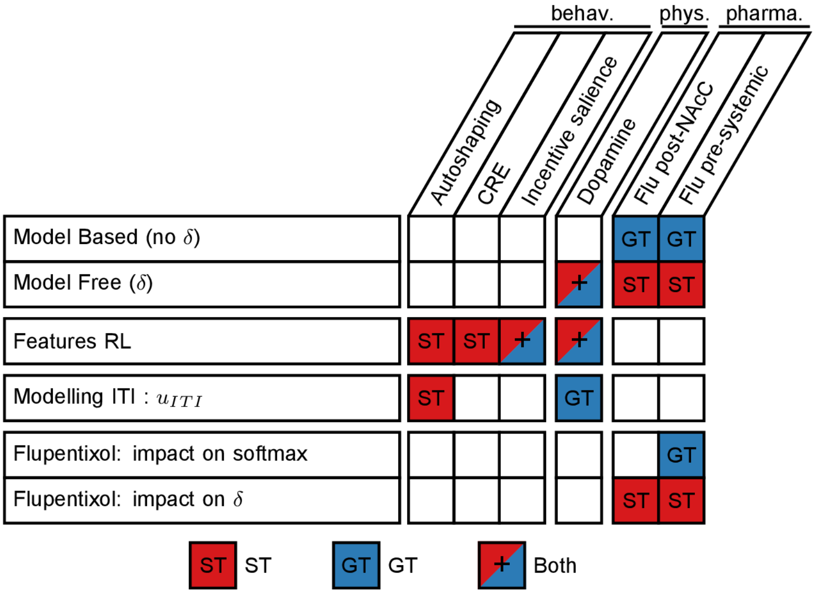
● yes × no

Figure 5.3: Summary of simulations and results. *Each line represents a different model composed of a pair of Reinforcement Learning systems. Each column represents a simulated experiment. Experiments are grouped by the kind of data accounted for: behavioural (autoshaping [Fla+09; Fla+11b], CRE [RF09], Incentive salience [MB09; DB12b]), physiological [Fla+09] and pharmacological (Flu post-NAcC [SR12], Flu pre-systemic [Fla+09]). Variant 4 (i.e. Model-based / Model-Free without features) is not included as it failed to even reproduce the autoshaping behavioural results and was not investigated further.*

three groups (see [Mey+12] for a more recently defined criterion). At lever appearance, rats that significantly increased their engagement towards it (top 30%) were classified as STs, whereas rats that almost never engaged with the lever (bottom 30%) were classified as GTs (these latter animals engaged the food magazine upon CS presentation). The remaining rats, engaging in both lever and magazine approach behaviours were defined as the Intermediate Group (IGs) (see Figure 5.5 A, B). STs and GTs acquired their respective CRs at a similar rate over days of training [RF09].

The current model is able to reproduce such results (see Figure 5.5 C, D). By running a simulation for each group of rats, using different parameters (mainly varying the ω parameter) the model reproduces the different tendencies to engage with the lever ($\omega = 0.499$), with the magazine ($\omega = 0.048$) or to fluctuate between the two ($\omega = 0.276$). A high ω strengthens the influence of the Feature-Model-Free system, which learns to associate a high motivational value to the lever CS, and a sign-tracking CR dominates. A low ω increases the influence of the Model-Based system, which infers the optimal behaviour to maximize reward, and goal-tracking is favoured. When both systems are mixed, i.e. with an intermediate ω, the behaviour is more likely to oscillate between sign- and goal-tracking, representative of the intermediate group.

These results rely on the combination of two systems that would independently lead to ‘pure’ sign-tracking or goal-tracking CRs. Three tested variants of the model could reproduce these behavioural results as well (see Figure S5.1): a combination of Feature-Model-Free systems and simple Model-Free system (Variant 1); a multi-step extension of Dayan 2006’s model [Day+06] giving a Pavlovian impetus for the lever (Variant 2); and



	behav.			phys.	pharma.	
	Autoshaping	CRE	Incentive salience	Dopamine	Flu post-NAcC	Flu pre-systemic
Model Based (no δ)					GT	GT
Model Free (δ)				+	ST	ST
Features RL	ST	ST	+	+		
Modelling ITI : u_{ITI}	ST			GT		
Flupentixol: impact on softmax						GT
Flupentixol: impact on δ					ST	ST

Figure 5.4: Summary of the key mechanisms required by the model to reproduce experimental results. *Each line represents a different mechanism of the model. Each column represents a simulated experiment. For each mechanism, it states in which experiment and for which behaviour – sign-tracking (red), goal-tracking (blue) or both (+) – it is required. Note however that all mechanisms and associated parameters have, to a certain extent, an impact on any presented results.*

a symmetrical version of this last model with two impetuses, one for the lever, and one for the magazine (Variant 3) (see Methods). Interestingly, a combination of Model-Based and classical Model-Free (not feature-based : Variant 4) fails in reproducing these results (see Figure S5.8). This is because both systems are proven to converge to the same values and both would favour pure goal-tracking, such that varying their contribution has no impact on the produced behaviours.

Thus, at this stage, we can conclude that several computational models based on dual learning systems can reproduce these behavioural results, given that the systems favour different behaviours (see Figure S5.1). However, Variants 1, 2 and 3 fail to reproduce other behavioural, pharmacological and physiological data characteristic of STs and GTs (see following sections).

Incentive salience

The results in Figure 5.5 only represent the probability of approach to either the lever-CS or the food magazine. Thus, they do not account for the specific ways rats engage and interact with the respective stimuli. In fact, if food is used as the US, rats are known to chew and bite the stimuli on which they are focusing [MB09; DB12b] (see Figure 5.6 A). Importantly, both STs and GTs express this consumption-like behaviour during the CS period, directed towards the lever or the food magazine, respectively. It has been

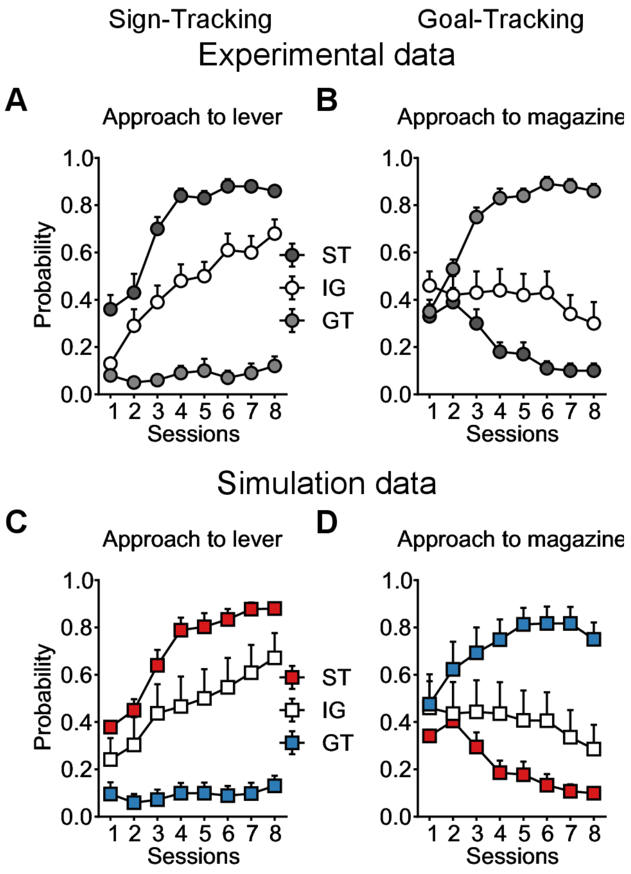


Figure 5.5: Reproduction of sign- versus goal-tracking tendencies in a population of rats undergoing an autoshaping experiment. *Mean probabilities to engage at least once with the lever* **(A,C)** *or the magazine* **(B,D)** *during trials. Data are expressed as mean ± S.E.M. and illustrated in 50-trial (2-session) blocks.* **(A,B)** *Reproduction of Flagel et al. [Fla+09] experimental results (Figure 2 A,B). Sign-trackers (ST) made the most lever presses (black), goal-trackers (GT) made the least lever presses (grey), Intermediate group (IG) is in between (white).* **(C,D)** *Simulation of the same procedure (squares) with the model. Simulated groups of rats are defined as STs (*$\omega = 0.499$*;* $\beta = 0.239$*;* $\alpha = 0.031$*;* $\gamma = 0.996$*;* $u_{ITI} = 0.027$ *;* $\mathcal{Q}_i(s_1, goL) = 0.844$*;* $\mathcal{Q}_i(s_1, exp) = 0.999$*;* $\mathcal{Q}_i(s_1, goM) = 0.538$*; n=14) in red, GTs (*$\omega = 0.048$ *;* $\beta = 0.084$*;* $\alpha = 0.895$*;* $\gamma = 0.727$*;* $u_{ITI} = 0.140$*;* $\mathcal{Q}_i(s_1, goL) = 1.0$*;* $\mathcal{Q}_i(s_1, exp) = 0.316$*;* $\mathcal{Q}_i(s_1, goM) = 0.023$*; n=14) in blue and IGs (*$\omega = 0.276$*;* $\beta = 0.142$*;* $\alpha = 0.217$*;* $\gamma = 0.999$*;* $u_{ITI} = 0.228$*;* $\mathcal{Q}_i(s_1, goL) = 0.526$*;* $\mathcal{Q}_i(s_1, exp) = 0.888$*;* $\mathcal{Q}_i(s_1, goM) = 0.587$*; n=14) in white. The model reproduces the same behavioural tendencies. With training, STs tend to engage more and more with the lever and less with the magazine, while GTs neglect the lever to increasingly engage with the magazine. IGs are in between.*

argued that this behaviour may reflect the degree to which incentive salience is attributed to these stimuli, and thus the extent to which they become "wanted" [MB09; DB12b; Ber07].

In an RL-like framework, incentive salience attribution can be represented as a bonus

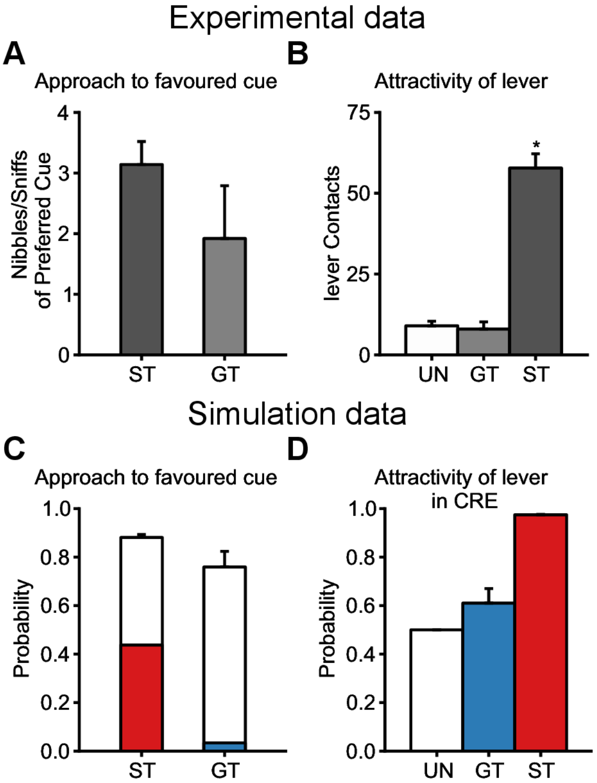


Figure 5.6: Possible explanation of incentive salience and Conditioned Reinforcement Effect by values learned during autoshaping procedure. *Data are expressed as mean ± S.E.M. Simulated groups of rats are defined as in Figure 5.5.* **(A)** *Number of nibbles and sniffs of preferred cue by STs and GTs as a measure for incentive salience. Data extracted from Mahler et al. [MB09] from Figure 3 (bottom-left).* **(B)** *Reproduction of Robinson et al. [RF09] experimental results (Figure 2 B). Lever contacts by STs and GTs during a conditioned reinforcer experiment.* **(C)** *Probability to engage with the respective favoured stimuli of STs and GTs at the end of the simulation (white, similar to the last session of Figure 5.5 C for STs and D for GTs) superimposed with the contribution in percentage of the values attributed by the Feature-Model-Free system in such engagement for STs (red) and GTs (blue). We hypothesize that such value is the source of incentive salience and explains why STs and GTs have a consumption-like behaviour towards their favoured stimulus.* **(D)** *Probability to engage with the lever versus exploring when presented with the lever and no magazine for STs (red), GTs (blue) and a random-policy group UN (white), simulating the unpaired group (UN) of the experimental data. Probabilities were computed by applying the softmax function after removing the values for the magazine interactions (see Methods). STs would hence actively seek to engage with the lever relatively to GTs in a Conditioned Reinforcement Effect procedure.*

mechanism for interacting with stimuli. The Feature-Model-Free system in the model realizes such a function, providing a specific bonus for each stimulus in any simulated rat. Such bonus was inspired by the Pavlovian impetus mechanism of Dayan 2006's model [Day+06]. Figure 5.6 C shows the percentage of Feature-Model-Free value that contributed to the computation of the probability to engage with the respective favoured cues of STs and GTs at the end of the simulation.

The presence of the magazine in the inter-trial interval (ITI), and the necessary revision of the associated bonus at a lower value when exploring, makes the associated bonus smaller than that of the lever (see Methods). This results in a even smaller contribution of this bonus in GTs behaviour (blue bar in Figure 5.6 C) compared to STs (red bar in Figure 5.6 C). Although it is not straightforward to interpret how the probability of engagement (white bars in Figure 5.6 C) in the model might be translated into a consumption-like behaviour from a computational point of view, we propose that the different contributions of bonuses could explain the slightly smaller number of nibbles and sniffs of preferred cue observed experimentally in GTs compared to STs (Figure 5.6 A, adapted from [MB09]). This may also explain why other studies have observed a smaller proportion of nibbles on the magazine in GTs [DB12b] and less impulsiveness [Lov+11] in GTs compared to STs. We come back to this issue in the discussion.

Variants 1 and 3 also realize such function by providing bonuses for actions leading to both stimuli (see Figure S5.2). Only providing bonus for sign-tracking behaviour – as in Dayan's model (Variant 2) – does not fit well with the attribution of incentive salience to both stimuli. It would suggest that we should not observe incentive salience towards the magazine in any rats, which is in discrepancy with the experimental data. Thus, the important mechanism here is that stimuli are not processed differently. Any stimulus is attributed with its respective bonus, which is pertinent in regard to the attribution of incentive salience.

Conditioned Reinforcement Effect (CRE)

An important question about the difference in observed behaviours is about the properties acquired by the lever that makes it more attractive to STs than to GTs. To answer this question, Robinson and Flagel studied the dissociation of the predictive and motivational properties of the lever [RF09]. Part of their results involves asking whether the Pavlovian lever-CS would serve as a conditioned reinforcer, capable of reinforcing the learning of a new instrumental response [Wil94b; Ski38]. In a new context, rats were presented with an active and an inactive nose port. Nose poking into the active port resulted in presentation of the lever for 2 seconds without subsequent reward delivery, whereas poking into the inactive one had no consequence. The authors observed that while both STs and GTs preferred the active nose port to an inactive one, STs made significantly more active nose pokes than GTs (see Figure 5.6 B, see also [Lom+11]). This suggests that the lever acquired greater motivational value in STs than in GTs.

Without requiring additional simulations, the model can explain these results by the value that has been incrementally learned and associated with approaching the lever in the prior autoshaping procedure for STs and GTs. In the model, STs attribute a higher value to interacting with the lever than GTs and should actively work for its appearance enabling further engagement. Figure 5.6 D shows the probabilities of engagement that would be computed at lever appearance after removing the magazine (and related actions) at the end of the experiment. Indeed, even though the lever is presented only very briefly, upon its presentation in the conditioned reinforcement test, STs actively engage and interact with it [RF09]. Any value associated to a state-action pair makes this action in the given state rewarding in itself, favouring actions (e.g. nosepokes) that would lead to such state. Repeatedly taking this action without receiving rewards should eventually lead to a decrease of this value and reduce the original engagement.

Physiological data

Not only have Flagel et al. [Fla+11b] provided behavioural data but they also provide physiological and pharmacological data. This raises the opportunity to challenge the model at different levels, as developed in the current and next sections.

Using Fast Scan Cyclic Voltammetry (FSCV) in the core of the nucleus accumbens they recorded the mean of phasic dopamine (DA) signals upon CS (lever) and US (food) presentation. It was observed that depending on the subgroup of rats, distinct dopamine release patterns emerge (see Figure 5.7 A,B) during Pavlovian training. STs display the classical propagation of a phasic dopamine burst from the US to the CS over days of training and the acquisition of conditioned responding (see Figure 5.7 A). This pattern of dopamine activity is similar to that seen in the firing of presumed dopamine cells in monkeys reported by Schultz and colleagues [Sch98] and interpreted as an RPE corresponding to the reinforcement signal δ of Model-Free RL systems [SB98]. In GTs, however, a different pattern was observed. Initially there were small responses to both the CS and US, of which the amplitudes seemed to follow a similar trend over training (see Figure 5.7 B).

By recording the mean of the RPEs δ computed in the Feature-Model-Free system during the autoshaping simulation (i.e. only fitted to behavioural data), the model can still qualitatively reproduce the different patterns observed in dopamine recordings for STs and GTs (see Figure 5.7 C,D). For STs, the model reproduces the progressive propagation of δ from the US to the CS (see Figure 5.7 C). For GTs, it reproduces the absence of such propagation. The RPE at the time of the US remains over training, while a δ also appears at the time of the CS (see Figure 5.7 D). In the model, such discrepancy is explained by the difference in the values that STs and GTs use for the computation of RPEs at the time of the CS and the US. STs, by repeatedly focusing on the lever, propagate the total value of food to the lever and end up having a unique δ at the unexpected lever appearance only. By contrast, by repeatedly focusing on the magazine during the lever appearance but, as all rats, also from time to time during ITI, GTs revise the magazine value multiple times, positively just after food delivery and negatively during ITI. Such revisions lead to a permanent discrepancy between the expected and observed value, i.e. a permanent δ, at lever appearance and food delivery, when engaging with the magazine.

The key mechanism to reproduce these results resides in the generalization capacities of the Feature-Model-Free system. Based on features rather than states, feature-values are to be used, and therefore revised, at different times and states of the experiment, favouring the appearance of RPEs. Variants 2, 3 and 4 relying on classical Model-Free systems are unable to reproduce such results (see Figure S5.3). By using values over abstract states rather than stimuli, it makes it impossible to only revise the value of the magazine during ITI. Therefore, given the deterministic nature of the MDP, we observe

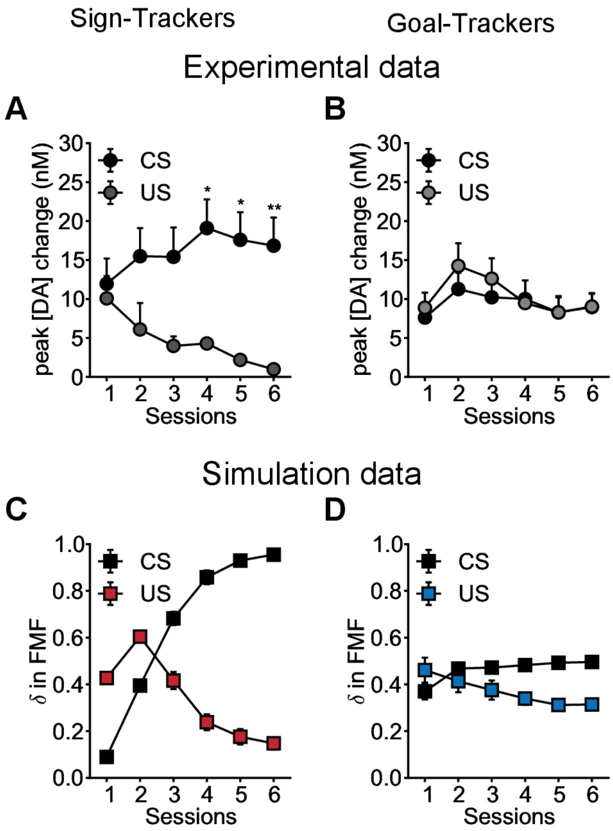


Figure 5.7: Reproduction of patterns of dopaminergic activity of sign- versus goal-trackers undergoing an autoshaping experiment. *Data are expressed as mean ± S.E.M.* **(A,B)** *Reproduction of Flagel et al. [Fla+11b] experimental results (Figure 3 d,f). Phasic dopamine release recorded in the core of the nucleus accumbens in STs (light grey) and GTs (grey) using Fast Scan Cyclic Voltammetry. Change in peak amplitude of the dopamine signal observed in response to CS and US presentation for each session of conditioning* **(C,D)** *Average RPE computed by the Feature-Model-Free system in response to CS and US presentation for each session of conditioning. Simulated groups of rats are defined as in Figure 5.5. The model is able to qualitatively reproduce the physiological data. STs (blue) show a shift of activity from US to CS time over training, while GTs develop a second activity at CS time while maintaining the initial activity at US time.*

a classical propagation of RPEs in all pathways up to the appearance of the lever.

Pharmacological data

Effects of systemic flupentixol administration on the learning of sign- and goal-tracking behaviours

Flagel et al. [Fla+11b] also studied the impact of systemic injections of the non specific dopamine antagonist, flupentixol, on the acquisition of sign-tracking and goal-tracking CRs. The authors injected flupentixol in rats prior to each of 7 sessions and observed the resulting behaviours. Behaviour during the 8th session was observed without flupentixol.

Systemic injections of flupentixol in STs and GTs (Flu groups, black curves in Figure 5.8 A,B) blocked expression of their respective behaviours during training. Saline injections (white curves in Figure 5.8 A,B) left their performances intact. The crucial test for learning took place on the 8th day, when all rats were tested without flupentixol. STs failed to approach the lever, and performed as the saline-injected controls did on the first day of training.

Thus, in STs flupentixol blocked the acquisition of a sign-tracking CR (see Figure 5.8 A). Interestingly, on the flupentixol-free test day GTs did not differ from the saline-injected control group, indicating that flupentixol did not block the acquisition of a goal-tracking CR (see Figure 5.8 B). Thus, acquisition of a sign-tracking CR, but not a goal-tracking CR, is dependent on dopamine (see also [DE10]).

The model reproduces these pharmacological results (see Figure 5.8 C,D). As in the experimental data, simulated GTs and STs do not show a specific conditioned response during the first 7 sessions under flupentixol. On the 8th session, without flupentixol, we observe that STs still do not show a specific conditioned response while GTs perform at a level close to that of the saline-injected control group (see Figure 5.8 C,D).

The absence of specific conditioned response in the whole population for the first 7 sessions is first due to the hypothesized [Hum+12] impact of flupentixol on action selection (see Methods). With enough flupentixol, the elevation of the selection temperature leads to a decrease of the influence of learned values in the expressed behaviour, masking any possibly acquired behaviour.

The absence of a specific conditioned response in STs is due to the blockade of learning in the second system by flupentixol, since it is RPE-dependent. Therefore almost no learning occurs in the system (see Figure 5.8).

In contrast, with the first system being RPE-independent, flupentixol has no effect on learning, because it is Model-Based rather than Model-Free [KH12]. The expression of behaviour is blocked at the action selection level, which does not make use of values learned by the Model-Based system. Thus, GTs, relying mainly on the first system, learn their CR under flupentixol but are just not able to express it until flupentixol is removed. The lower level of goal-tracking in the Flu group relative to the saline-injected control group on the 8th session is due to the lack of exploitation induced by flupentixol injection during the previous 7 sessions. By engaging less with the magazine, the Flu group ends up associating a lower value to the magazine (i.e. the value did not fully converge in 7

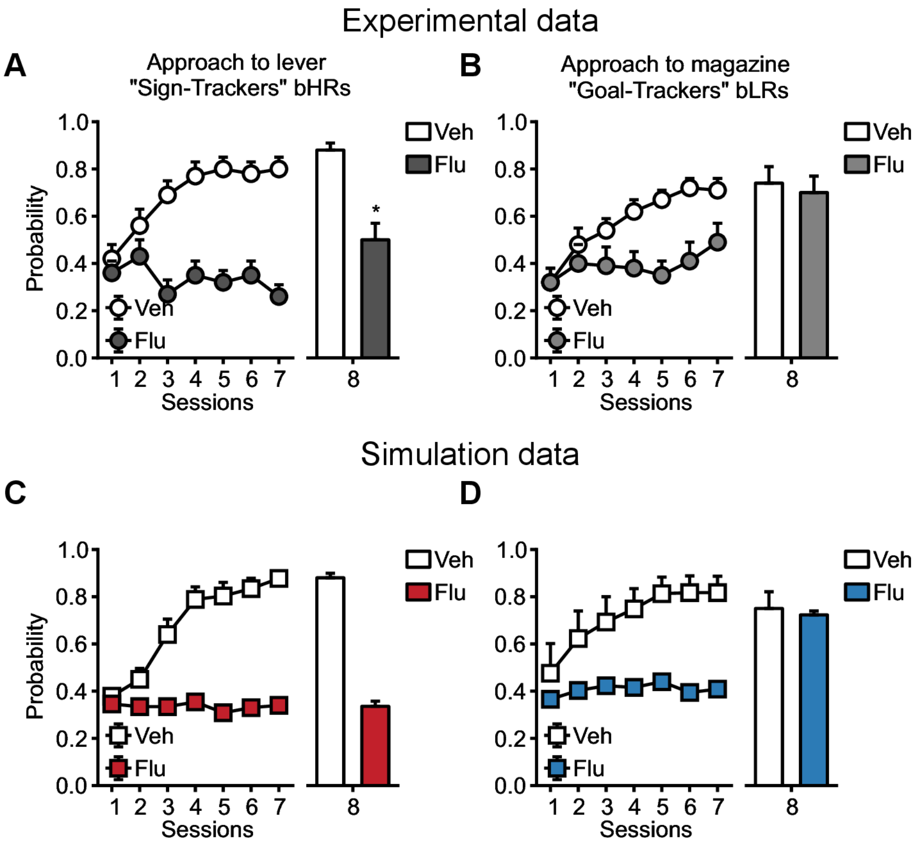


Figure 5.8: Reproduction of the effect of systemic injections of flupentixol on sign-tracking and goal-tracking behaviours. *Data are expressed as mean ± S.E.M.* **(A,B)** *Reproduction of Flagel et al. [Fla+11b] experimental results (Figure 4 a,d). Effects of flupentixol on the probability to approach the lever for STs* **(A)** *and the magazine for GTs* **(B)** *during lever presentation.* **(C,D)** *Simulation of the same procedure (squares) with the model. Simulated groups of rats are defined as in Figure 5.5.* **(C)** *By flattening the softmax temperature and reducing the RPEs of the Feature-Model-Free system, to mimic the possible effect of flupentixol, the model can reproduce the blocked acquisition of sign-tracking in STs (red), engaging less the lever relatively to a saline-injected control group (white).* **(D)** *Similarly, the model reproduces that goal-tracking was learned but its expression was blocked. Under flupentixol (first 7 sessions), GTs (blue) did not express goal-tracking, but on a flupentixol-free control test (8th session) their engagement with the magazine was almost identical to the engagement of a saline-injected control group (white).*

sessions) to guide its behaviour.

Interestingly, if the model had been constituted of Model-Free systems only – as in Variants 1, 2 and 3 – it would not have been able to reproduce these results, because both systems would have been RPE-dependent and thus sensitive to the effect of flupentixol (see Figure S5.4).

Effects of local flupentixol administration on the expression of sign- and goal-tracking behaviours

In a related experiment, Saunders et al. [SR12] studied the role of dopamine in the nucleus accumbens core in the expression of Pavlovian-conditioned responses that had already been acquired. After the same autoshaping procedure as in [Fla+07], they injected different doses of flupentixol in the core of the nucleus accumbens of rats and quantified its impact on the expression of sign-tracking and goal-tracking CRs in an overall population (without distinguishing between STs and GTs).

They found that flupentixol dose dependently attenuated the expression of sign-tracking, while having essentially no effect on goal-tracking (see Figure 5.9 A, B). Along with the Flagel et al. [Fla+11b] study, these results suggest that both the acquisition and expression of a sign-tracking CR is dopamine-dependent (at least in the core) whereas the acquisition and expression of a goal-tracking CR is not.

Given the assumption that the Feature-Model-Free system would take place in or rely on the core of the nucleus accumbens, this model reproduces the main experimental result: the decreased tendency to sign-track in the population (see Figure 5.9 C). Note that in the previous experiment, the injection of flupentixol was systemic, and assumed to affect any region of the brain relying on dopamine, whereas in the present experiment it was local to the core of the nucleus accumbens. Therefore, we modelled the impact of flupentixol differently between the current and previous simulations (see Methods). In the model, the tendency to sign-track is directly correlated with a second operational system. Any dysfunction in the learning process (here by a distortion of RPEs) reduces this trend.

The model successfully reproduced the absence of reduction of goal-tracking, in contrast to the reduction of sign-tracking. However, it was unable to reproduce the invariance in goal-tracking (see Figure 5.9 D) and rather produced an increase in goal-tracking. This is due to the use of a softmax operator for action selection, as this is the case in the vast majority of computational neuroscience RL models [Day+06; Glä+10; Hum+12; Daw+05; Ker+11; Huy+12; Doy+02; Red+07], which automatically favours goal-tracking when sign-tracking is blocked (see Limitations). We did not attempt to cope with this limitation because our focus here was the absence of reduction of goal-tracking.

Besides, the model could, after re-learning, reproduce the selective impact of intra-accumbal flupentixol injections observed in sign-tracking but not in goal-tracking, because such injections affected the learning process in the Feature-Model-Free system only.

Discussion

We tested several mechanisms from the current literature on modelling individual variation in the form of Pavlovian conditioned responses (ST vs GT) that emerge using a classical autoshaping procedure, and the role of dopamine in both the acquisition and expression of these CRs. Benefiting from a rich set of data, we identified key mechanisms that are sufficient to account for specific properties of the observed behaviours. The resulting model relies on two major concepts: Dual learning systems and factored

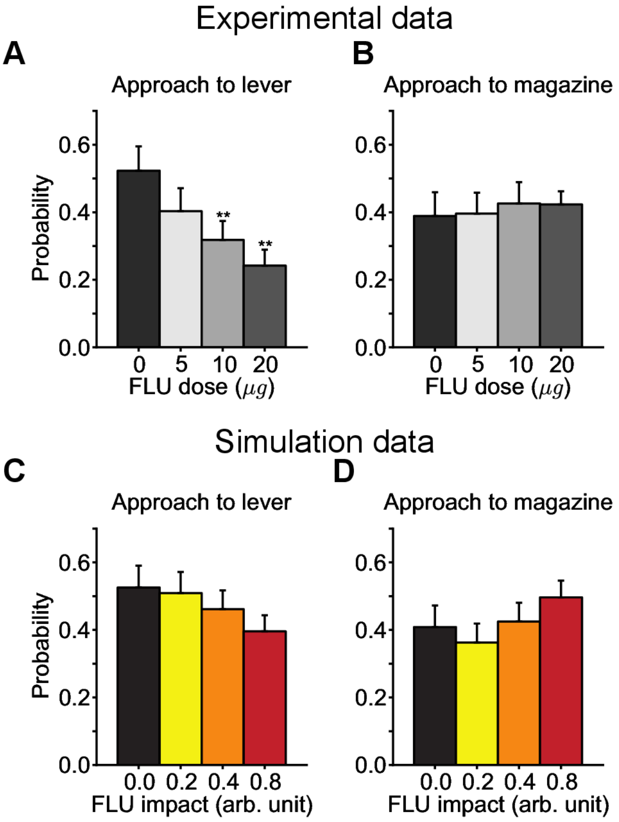


Figure 5.9: Reproduction of the effect of post injections of flupentixol in the core of the nucleus accumbens. *Data are expressed as mean ± S.E.M.* **(A,B)** *Reproduction of Saunders et al. [SR12] experimental results (Figure 2 A,D). Effects of different doses of flupentixol on the general tendency to sign-track* **(A)** *and goal-track* **(B)** *in a population of rats, without discriminating between sign- and goal-trackers.* **(C,D)** *Simulation of the same procedure with the model. The simulated population is composed of groups of rats defined as in Figure 5.5. By simulating the effect of flupentixol as in Figure 5.8, the model is able to reproduce the decreasing tendency to sign-track in the overall population by increasing the dose of flupentixol.*

representations. Figure 5.4 summarizes the role of each mechanism in the model.

Dual learning systems

Combining Model-Based and Model-Free systems has previously been successful in explaining the shift from goal-directed to habitual behaviours observed in instrumental conditioning [Daw+05; Ker+11; Huy+12; KH12; Glä+10]. However, few models based on the same concept have been developed to account for Pavlovian conditioning [Day+06]. While the need for two systems is relevant in instrumental conditioning given the distinct temporal engagement of each system, such a distinction has not been applied to Pavlovian phenomena (but see recent studies on orbitofrontal cortex [Tak+09; McD+11; McD+12]). The variability of behaviours and the need for multiple systems have been masked by focusing on whole populations and, for the most part, ignoring individual differences in studies of Pavlovian conditioning. The nature of the CS is especially important, as many studies of Pavlovian conditioned approach behaviour have used an auditory stimulus as the CS, and in such cases only a goal-tracking CR emerges in rats [CD83; Mey+10].

As expected from the behavioural data, combining two learning systems was successful in reproducing sign- and goal-tracking behaviours. The Model-Based system, learning the structure of the task, favours systematic approach towards the food magazine, and waiting for food to be delivered, and hence the development of a goal-tracking CR. The Feature-Model-Free system, directly evaluating features by trials and errors, favours systematic approach towards the lever, a full predictor of food delivery, and hence the development of a sign-tracking CR. Moreover, utilizing the Feature-Model-Free system to represent sign-tracking behaviour yields results consistent with the pharmacological data. Disrupting RPEs, which reflects the effects of flupentixol on dopamine, blocks the acquisition of a sign-tracking CR, but not a goal-tracking CR. The model does not make a distinction between simple approach behaviour versus consumption-like engagement, as reported for both STs and GTs [MB09; DB12b]. However given that such engagement results from the development of incentive salience [MB09; DB12b], the values learned by the Feature-Model-Free system to bias behaviour towards stimuli attributed with motivational value are well-suited to explain such observations. The higher motivational value attributed to the lever by STs relative to GTs can also explain why the lever-CS is a more effective conditioned reinforcer for STs than for GTs [RF09].

Importantly, none of the systems are dedicated to a specific behaviour, nor rely on *a priori* information to guide their processes. The underlying mechanisms increasingly make one behaviour more pronounced than the other through learning. Each system contributes to a certain extent to sign- and goal-tracking behaviour. This property is emphasized by the weighted sum integration of the values computed by each system before applying the softmax action-selection mechanism. The variability of behaviours in the population can then be accounted for by adjusting the weighting parameter ω from 1 (i.e. favouring sign-tracking) to 0 (i.e. favouring goal-tracking). This suggests that the rats' actions result from some combination of rational and impulsive processes, with individual variation contributing to the weight of each component.

The integration mechanism is directly inspired by the work of Dayan et al. [Day+06] and as the authors suggest, the parameter ω may fluctuate over time, making the contribution of the two systems vary with experience. In contrast to their model, however, the model presented here does not assign different goals to each system. Thus, the current model is more similar to their previous model [Daw+05], which uses another method for integration.

A common alternative to integration when using multiple systems [Daw+05; Ker+11; Doy+02] is to select at each step, based on a given criterion (certainty, speed/accuracy trade-off, energy cost), a single system to pick the next action. Such switch mechanism does not fit well with the present model, given that it would be interpreted as if actions relied sometimes only on motivational values (i.e. Feature-Model-Free system) and sometimes only on a rational analysis of the situation (i.e. Model-Based system). It also does not fit well with pharmacological observation that STs do not express goal-tracking tendencies in the drug-free test session following systemic-injections of flupentixol [Fla+11b], as Flagel et al. stated, "[sign-tracking] rats treated with flupentixol did not develop a goal-tracking CR".

Factored representations

Classical RL algorithms used in neuroscience [Daw+05; Ker+11; Doy+02; Day+06], designed mainly to account for instrumental conditioning, work at the state level. Tasks are defined as graphs of states, and corresponding models are unaware of any similarity within states. Therefore, any subsequent valuation process cannot use any underlying structure to generalize updates to states that share stimuli. Revising the valuation process to handle features rather than states *per se*, makes it possible to attribute motivational values to stimuli independently of the states in which they are presented.

Recent models dedicated to Pavlovian conditioning [Sch+96; Bal99; Red+07; SM07; Cou+06; GN12] usually represent and process stimuli independently and can be said to use factored representations, a useful property to account for phenomena such as blocking [Kam67] or overexpectation [LN98]. In contrast to the present model, while taking inspiration from RL theory (e.g. using incremental updates), these models are usually far from the classical RL framework. Of significant difference with the present study, most of these models tend to describe the varying intensity of a unique conditioned response and do not account for variations in the actual form of the response, as we do here. In such models, the magazine would not be taken into account and/or taken as part of the context, making it unable to acquire a value for itself nor be the focus of a particular response.

In RL theory, factorization is mainly evoked when trying to overcome the curse of dimensionality [Bel57] (i.e. standard algorithms do not scale well to high dimensional spaces and require too much physical space or computation time). Amongst methods that intend to overcome this problem are value function approximations and Factored Reinforcement Learning. Value function approximations [Doy+02; Kha+06; Elf+13] attempt to split problems into orthogonal subproblems making computations easier and providing valuations that can then be aggregated to estimate the value of states. Factored Reinforcement Learning [Bou+00; Deg+06; VB08] attempts to find similarities between states so that they can share values, reducing the physical space needed and relies on factored Markov Decision Processes. We also use factored Markov Decision processes, hence the "factored" terminology. However, our use of factored representations serves a different purpose. We do not intend to build a compact value-function nor infer the value of states from values of features but rather make these values compete in the choice for the next action.

Taking advantage of factored representations into classical RL algorithms is at the very heart of the present results. By individually processing stimuli within states (i.e. in the same context, at the same time and same location) and making them compete, the Feature-Model-Free system favours a different policy – oriented towards engaging with the most valued stimuli – (sign-tracking) than would have been favoured by classical algorithms such as Model-Based or Model-Free systems (goal-tracking). Hence, combining a classical RL algorithm with the Feature-Model-Free system enables the model to reproduce the difference in behaviours observed between STs and GTs during an autoshaping procedure. Moreover, by biasing expected optimal behaviours towards cues with motivational values (incentive salience), it is well suited to explain the observed commitment to unnecessary and possibly counter-productive actions (see also [Day+06; GM+12; Huy+11]). Most of all, it enables the model to replicate the different patterns of dopamine activity recorded with FSCV in the core of the nucleus accumbens of STs and GTs. The independent processing of stimuli leads to patterns of RPE that match those of dopamine activity for STs – a shift of bursts from the US to the CS; and in GTs – a persistence of bursts at both the time of the US and the CS.

A promising combination

By combining the two concepts of dual learning systems and factored representations in a single model, we are able to reproduce individual variation in behavioural, physiological and pharmacological effects in rats trained using an autoshaping procedure. Interestingly, our approach does not require a deep revision of mechanisms that are extensively used in our current field of research.

While Pavlovian and instrumental conditioning seem entangled in the brain [Yin+08], the two major concepts on which rely their respective models, dual learning systems and factored representations, have to our knowledge never been combined into a single model in this field of research.

This approach could contribute to the understanding of interactions between these two classes of learning, such as CRE or Pavlovian-Instrumental Transfer (PIT), where motivation for stimuli acquired via Pavlovian learning modulates the expression of instrumental responses. Interestingly, the Feature-Model-Free system nicely fits with what would be expected from a mechanism contributing to general PIT [CB05]. It is focused on values over stimuli without regard to their nature [CB05], it biases and interferes with some more instrumental processes [CB05; Huy+11; GM+12] and it is hypothesized to be located in the core of the nucleus accumbens [CB05]. It would thus be interesting to study whether future simulations of the model could explain and help better formalize these aspects of PIT.

We do not necessarily imply that instrumental and Pavlovian conditioning might rely on a unique model. Rather, we propose that if they were the results of separated systems, they should somehow rely on similar representations and valuation mechanisms, given the strength of the observed interactions.

Theoretical and practical implications

The proposed model explains the persistent dopamine response to the US in GTs over days of training as a permanent RPE due to the revision of the magazine value during each ITI. Therefore, a prediction of the model is that shortening the ITI should reduce the amplitude of this burst (i.e. there should be less time to revise the value and reduce the size of the RPE); whereas increasing the ITI should increase the amplitude of this burst. Removing the food dispenser during ITI, similar to theoretically suppressing the ITI, should make this same burst disappear. Studying physiological data by grouping them given the duration of the preceding ITI might be sufficient, relatively to noise, to confirm that its duration impacts the amplitude of dopamine bursts. In the current experimental procedure, the ITI is indeed randomly picked in a list of values with an average of 90 sec. Moreover, reducing ITI duration should lead to an increase of the tendency to goal-track in the overall population. Indeed, with a higher value of the food magazine, the Feature-Model-Free system would be less likely to favour sign-tracking over goal-tracking CR. The resulting decrease in sign-tracking in the overall population would be consistent with findings of previous works [BP79; GB81; GG00; Tom+03], where a shorter ITI reduces the observed performance in the acquisition of sign-tracking CRs. Alternatively, it would also be interesting to examine the amplitude of dopamine bursts during the ITI (especially when exploring the food magazine), to determine whether or not physiological responses during this period affect the outcome of the conditioned response.

It would be interesting to split physiological data not only between STs and GTs but also between the stimuli on which the rats started and/or ended focusing on during CS presentation at each trial. This would help to confirm that the pattern of dopamine activity is indeed due to a separate valuation of each stimuli. We would predict that at the time of the US, dopamine bursts during engagement with the lever should be small relatively to dopamine bursts during engagement with the magazine. Moreover, comparing dopamine activity at the time of the CS when engaging with the lever versus the magazine could help elucidate which update mechanism is being used. If activity differs, this would suggest that the model should be revised to use SARSA-like updates, i.e. taking into account the next action in RPE computation. Such a question has already been the focus of some studies on dopamine activity [Mor+06; Roe+07; Bel+12b].

There is no available experimental data for the phasic dopaminergic activity of the intermediate group. The model predicts that such a group would have a permanent phasic dopamine burst, i.e. RPE, at US and a progressively appearing burst at CS (see Figure S5.6). Over training, the amplitude of the phasic dopamine burst at US should decrease until a point of convergence, while at the mean time the response at CS should increase until reaching a level higher than the one observed at US. However, one must note, that the fitting of the intermediate group is not as good as for STs or GTs, as it regroups behaviours that range from sign-tracking to goal-tracking, such that this is a weak prediction.

There is the possibility that regularly presenting the magazine or the lever could, without pairing with food, lead to responses that are indistinguishable from CRs. However, ample evidence suggests that the development of a sign-tracking or goal-tracking CR is not due to this pseudoconditioning phenomenon, but rather a result of learned CS-US associations. That is, experience with lever-CS presentations or with food US does not account for the acquisition of lever-CS induced directed responding [Tom+12; RF09]. Nonetheless, it should be noted that the current model cannot distinguish between pseudoconditioning CR-like responses and sign-tracking or goal-tracking behaviours. This would require us to introduce more complex MDPs that embed the ITI and can more clearly distinguish between approach and engagement.

Limitations

The Feature-Model-Free system presented in this article was designed as a proof of concept for the use of factored representations in computational neuroscience. In its present form it updates the value of one feature (the focused one) at a time, and this is sufficient to account for much of the experimental data. It does not address whether multiple features could be processed in parallel, such that multiple synchronized, but independently computed, signals would update distinct values relative to the attention paid to the associated features. Further experiments should be performed to confirm this hypothesis. Subsequently, using factored representations in the Model-Based system was not necessary to account for the experimental data and the question remains whether explaining some phenomena would require it.

While using factored representations, our approach still relies on the discrete-time state paradigm of classical RL, where updates are made at regular intervals. Although such simplification can explain the set of data considered here, one would need to extend this to continuous time if one would like to also model experimental data where rats take more or less time to initiate actions that can vary in duration [Fla+11b]. The present model, which does not take timing into consideration, cannot account for the fact that STs and GTs both come to approach their preferred stimuli faster and faster as a function of training nor does it make use of the variations of ITI duration. Our attempt to overcome this limitation using the MDP framework was unsuccessful. Focusing on features, it becomes more tempting to deal with the timing of their presence, a property that is known to be learned and to have some impact on behaviours [GG00; KS08; Daw+06c; Fio+08].

Moreover, in the current model, we did not attempt to account for the conditioned orienting responses (i.e. orientation towards the CS) that both STs and GTs exhibit upon CS presentation [SR12]. However, we hypothesize that such learned orienting responses could be due to state discrimination mechanisms that are not included in the model, and would be better explained with partial observability and actions dedicated to collect information. This is beyond the scope of the current article, but is of interest for future studies.

As evident by the only partial reproduction of the flupentixol effects on the expression of sign- and goal-tracking behaviours, the model is limited by the use of the softmax action-selection mechanism, which is widely used in computational neuroscience [Day+06; Glä+10; Hum+12; Daw+05; Ker+11; Huy+12; Doy+02; Red+07]. In the model, all actions are equal – there is no action with a specific treatment – and the action-selection mechanism necessarily selects an action at each time step. Any reduction in the value of one action favours the selection of all other actions in proportion to their current associated values. In reality, however, blocking the expression of an action would certainly lead mainly to inactivity rather than necessarily picking the alternative and almost never expressed action. One way of improving the model in this direction could be to replace the classical softmax function by a more realistic model of action selection in the basal ganglia (e.g. [Gur+04]). In such a model, no action is performed when no output activity gets above a certain threshold. Humphries et al. [Hum+12] have shown that changing the exploration level in a softmax function can be equivalent to changing the level of tonic dopamine in the basal ganglia model of Gurney et al. [Gur+04]. Interestingly, in the latter model, reducing the level of tonic dopamine results in difficulty in initiating actions and thus produces lower motor behaviour, as is seen in Parkinsonian patients and as can be seen in rats treated with higher doses of flupentixol [Fla+11b]. Thus a natural sequel to the current model would be to combine it with a more realistic basal ganglia model for action selection.

We simulated the effect of flupentixol as a reduction of the RPE in the learning processes of Model-Free systems to parallel its blockade of the dopamine receptors. While this is sufficient to account for the pharmacological results previously reported [Fla+11b], it fails to account for some specific aspects that have more recently emerged. Mainly, it is unable to reproduce the instant decreased engagement observed at the very first trial after post-training local injections of flupentixol [SR12]. Our current approach requires re-learning to see any impact of flupentixol. A better understanding of the mechanisms that enable instant shifts in motivational values, by shifts in the motivational state [RB13] or the use of drugs [SR12; Fla+11b], might be useful to extend the model on such aspects.

We also tried to model the effect of flupentixol on RPEs with a multiplicative effect, as it would have accounted for an instant impact on behaviour. However, it failed to account for the effects of flupentixol on learning of the sign-tracking CRs, as a multiplicative effect only slowed down learning but did not disrupt it. How to model the impact of flupentixol, and dopamine antagonists or drugs such as cocaine remains an open question (e.g. see [Pan+07; Red04]).

Finally, our work does not currently address the anatomical counterpart of ω at the heart of the model, nor the regions of the brain that would match the current Model-Based system and the Feature-Model-Free system. Numerous studies have already discussed the potential substrates of Model-Based / Model-Free systems in the prefrontal cortex / dorsolateral striatum [Daw+06a], or the dorsomedial and dorsolateral striatum [YK06; Tho+10; BD11; KH12; Mee+12]. The weighted sum integration may suggest a crossed projection of brains regions favouring sign- and goal-tracking behaviours (Model-Based and Feature-Model-Free systems) into a third one. We postulate there is a difference in strength of “connectivity” between such regions in STs vs GTs [Fla+11a]. Further, one might hypothesize that the core of the nucleus accumbens contributes to the Feature-Model-Free system. The integration and action selection mechanisms would naturally fit within the basal ganglia, stated to contribute to such functions [Min96; Red+99b; Gur+01; Hum+12].

Conclusion

Here we have presented a model that accounts for variations in the form of Pavlovian conditioned approach behaviour seen during autoshaping in rats; that is, the development of a sign-tracking vs goal-tracking CR. This works adds to an emerging set of studies suggesting the presence and collaboration of multiple RL systems in the brain. It questions the classical paradigm of state representation and suggests that further investigation of factored representations in RL models of Pavlovian and instrumental conditioning experiments may be useful.

Methods

Modelling the autoshaping experiment

In the classical reinforcement learning theory [SB98], tasks are usually described as Markov Decision Processes (MDPs). As the proposed model is based on RL algorithms, we use the MDP formalism to computationally describe the Pavlovian autoshaping procedure used in all simulations.

An MDP describes the interactions of an agent with its environment and the rewards it might receive. An agent being in a state s can execute an action a which results in a new state s' and the possible retrieval of some reward r. More precisely, an agent can be in a finite set of states S, in which it can perform a finite set of discrete actions A, the consequences of which are defined by a transition function $\mathcal{T} : S \times A \rightarrow \Pi(S)$, where $\Pi(S)$ is the probability distribution $\mathcal{P}(s'|s,a)$ of reaching state s' doing action a in state s. Additionally, the reward function $\mathcal{R} : S \times A \rightarrow \mathbb{R}$ is the reward $\mathcal{R}(s,a)$ for doing action a in state s. Importantly, MDPs should theoretically comply with the Markov property: the probability of reaching state s' should only depend on the last state s and the last action a. An MDP is defined as episodic if it includes at least one state which terminates the current episode.

Figure 5.1 shows the deterministic MDP used to simulate the autoshaping procedure. Given the variable time schedule (30-150s) and the net difference observed in behaviours in inter-trial intervals, we can reasonably assume that each experimental trial can be simulated with a finite horizon episode.

The agent starts from an empty state (s_0) where there is nothing to do but explore. At some point the lever appears (s_1) and the agent must make a critical choice: It can either go to the lever (s_2) and engage with it (s_5), go to the magazine (s_4) and engage with it (s_7) or just keep exploring (s_3,s_6). At some point, the lever is retracted and food is delivered. If the agent is far from the magazine (s_5,s_7), it first needs to get closer. Once close (s_7), it consumes the food. It ends in an empty state (s_0) which symbolizes the start of the inter-trial interval (ITI): no food, no lever and *an empty but still present magazine.*

The MDP in Figure 5.1 is common to all of the simulations and independent of the reinforcement learning systems we use. STs should favour the red path, while GTs should favour the *shorter* blue path. All of the results rely mainly on the action taken at the lever appearance (s_1), when choosing to go to either the lever, the magazine, or to explore. Exploring can be understood as not going to the lever nor to the magazine.

To fit with the requirements of the MDP framework, we introduce two limitations in our description, which also simplify our analyses. We assume that engagement is necessarily exclusive to one or no stimulus, and we make no use of the precise timing of the procedure – the ITI duration nor the CS duration – in our simulations.

Inter-trial interval (ITI)

While the MDP does not model the ITI, the results regarding physiological data rely partially on its presence. Extending the MDP with a set of states to represent this interval would increase the complexity of the MDP and the time required for simulations. The behaviour that could have resulted from such an extension is easily replaced by applying the following formula at the beginning of each episode:

$$\mathcal{V}(M) \leftarrow (1 - u_{ITI}) \times \mathcal{V}(M) \tag{5.1}$$

where the parameter $0 \leq u_{ITI} \leq 1$ reflects the interaction with the magazine that occurred during the ITI. A low $u_{ITI} \rightarrow 0$ symbolizes a low interaction and therefore a low revision of the value associated to the magazine. A high $u_{ITI} \rightarrow 1$ symbolizes a strong exploration of the magazine during the inter-trial interval and therefore a strong decrease in the associated value due to unrewarded exploration.

Model

The model relies on the architecture shown in Figure 5.2. The main idea is to combine the computations of two distinct reinforcement learning systems to define what behavioural response is chosen at each step.

Model-Based system (MB)

The first system is Model-Based [SB98], and classically relies on a transition function $\mathcal{T}$ and a reward function $\mathcal{R}$ which are learned by experience given the following rules:

$$\mathcal{T}(s,a,s') \leftarrow \begin{cases} (1-\alpha) \times \mathcal{T}(s,a,s'') + \alpha & \text{if } s' = s'' \\ (1-\alpha) \times \mathcal{T}(s,a,s'') & \text{otherwise} \end{cases} \tag{5.2}$$

$$\mathcal{R}(s,a) \leftarrow \mathcal{R}(s,a) + \alpha(r - \mathcal{R}(s,a)) \tag{5.3}$$

where the learning rate $0 \leq \alpha \leq 1$ classically represents the speed at which new experiences replace old ones. Using a learning rate rather than counting occurrences is a requirement for accordance with the incremental expression of the observed behaviours. This can account for some resistance or uncertainty in learning from new experiences.

Given this model, an action-value function $\mathcal{Q}$ can then be computed with the following classical formula:

$$\mathcal{Q}(s,a) \leftarrow \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \max_{a'} \mathcal{Q}(s',a') \tag{5.4}$$

where the discount rate $0 \leq \gamma \leq 1$ classically represents the preference for immediate versus distant rewards. The resulting Advantage function $\mathcal{A}$ [Bai93; DB02], the output of the first system, is computed as follows:

$$\mathcal{A}(s,a) \leftarrow \mathcal{Q}(s,a) - \max_{a'} \mathcal{Q}(s,a') \tag{5.5}$$

It defines the (negative) advantage of taking action a in state s relatively to the optimal action known. The optimal action therefore has an advantage value of 0.

In terms of computation, the advantage function could be replaced by the action-value function without changing the simulation results (we only compare $\mathcal{A}$-values over the same state and therefore $\max_{a'} \mathcal{Q}(s, a')$ is constant whatever the action). It has been used in preceding works dealing with interactions between instrumental and Pavlovian conditioning [DB02; Day+06] and we kept it for a better and more straightforward comparison with variants of the model that were directly inspired by these preceding works.

Feature-Model-Free system (FMF)

A state is generally described by multiple features. Animals, especially engaged in a repetitive task, might not pay attention to all of them at once. For example, when the lever appears and a rat decides to engage with the magazine, it focuses primarily on the magazine while ignoring the lever, such that it could update a value associated to the magazine but leave intact any value related to the lever (see Figure 5.10 A). Although this could be related to an attentional process that bias learning, we do not pretend to model attention with such a mechanism.

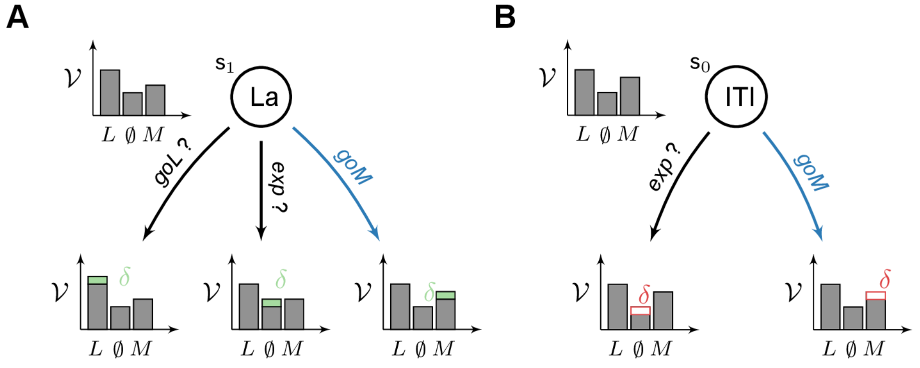


Figure 5.10: Characteristics of the Feature-Model-Free system. (A) *Focusing on a particular feature. The Feature-Model-Free system relies on a value function $\mathcal{V}$ based on features. Choosing an action (e.g.* goL, goM *or* exp*), defines the feature it is focusing on (e.g.* L*ever,* M*agazine or nothing ∅). Once the action is chosen (e.g.* goM *in blue), only the value of the focused feature (e.g. $\mathcal{V}(M)$) is updated by a standard reward prediction error, while leaving the values of the other features unchanged.* **(B)** *Feature-values permit generalization. At a different place and time in the episode, the agent can choose an action (e.g.* goM *in blue) focusing on a feature (e.g.* M*) that might have already been focused on. This leads to the revision of the same value (e.g. $\mathcal{V}(M)$) for two different states (e.g. s_1 and s_0). Values of features are shared amongst multiple states.*

Relying on this idea, the second system is a revision of classical Model-Free systems which is based on features rather than states. It relies on a value function $\mathcal{V} : \mathcal{C} \rightarrow \mathbb{R}$ based on a set of features $\mathcal{C}$, which is updated with an RPE:

$$\begin{aligned}\mathcal{V}(c(s,a)) &\leftarrow \mathcal{V}(c(s,a)) + \alpha\delta \\ \delta &\leftarrow r + \gamma \max_{a'} \mathcal{V}(c(s',a')) - \mathcal{V}(c(s,a))\end{aligned} \tag{5.6}$$

where $c : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{C}$ is a feature-function that returns the feature $c(s,a)$ the action a was focusing on in state s (see Table S5.2; Figure 5.1 also embeds the features returned by c for each action and state). One could argue that this feature-function, defined *a priori*, introduces an additional requirement relative to classical Model-Free systems. This is a weak requirement since this function is straightforward when actions, instead of being abstractly defined, are described as interactions towards objects in the environment. This function simply states that, for example, when pressing a lever, the animal is focusing on the lever rather than on the magazine. Similar to $\mathcal{Q}$-learning, we assume that the future action to be chosen is the most rewarding one. Therefore, the value chosen for the reached state s', in the computation of the RPE, is the highest value reachable by any possible future action $\max_{a'} \mathcal{V}(c(s',a'))$.

Classical Model-Free systems do not permit generalization in their standard form: even when two states share most of their features, updating the value of one state leaves the value of the other untouched. This new system overcomes such limitation (see Figure 5.10 B). In Feature-Model-Free Reinforcement Learning, multiple states in time and space can share features and their associated values. For example, while in ITI, rats tend from time to time to explore the magazine [RF09; Mey+12], which might lead them to revise any associated value, which can also be used when the lever appears. Therefore, actions in ITIs might impact the rest of the experiment.

In the simulated experiment (see Figure 5.1), this generalization phenomenon happens as follows: Assuming that the simulated rat was engaging the magazine (eng) before food delivery (from s_4 to s_7), then the value $\mathcal{V}$ of $c(s_4, \text{eng}) = M$ is updated with the following $\delta = 0 + \gamma \max'_a \mathcal{V}(c(s_7, a')) - \mathcal{V}(M)$. As the best subsequent action (and, for simplification, the only possible one) is to consume the food (in s_7), it results in a positive $\delta = \gamma\mathcal{V}(F) - \mathcal{V}(M)$. During ITI (which in the MDP is simulated by the u_{ITI} parameter), if the simulated rat checks the magazine (goM) and finds no food, then $\mathcal{V}(M)$ is revised with a negative $\delta = \gamma\mathcal{V}(\emptyset) - \mathcal{V}(M)$ (Figure 5.10 B). The value $\mathcal{V}(M)$ is therefore revised at multiple times in the experiment and, for example, a decrease of value during ITI has an impact on the choice of engaging with the magazine (goM) at lever appearance.

Processing features rather than states and the generalization that results from it is a key mechanism of the presented model. It makes the system favour a different path than the one favoured by classical reinforcement learning systems.

Contrary to what the system suggests, it is almost certain that rats might handle multiple features at once and could simultaneously update multiple values. We present here a version without such capacity since it is not required in the simulated experiments and simplifies its understanding.

Integration

The Feature-Model-Free system accounts for motivational bonuses $\mathcal{V}$ that impact values $\mathcal{A}$ computed by the Model-Based system. The integration of these values is made through a weighted sum:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}(s,a) + \omega\mathcal{V}(c(s,a)) \tag{5.7}$$

where $0 \leq \omega \leq 1$ is a combination parameter which defines the importance of each system in the overall model. ω is equivalent to the responsibility signal in Mixture of Experts [Jac+91; Doy+02]. We want to emphasize that the two systems are not in simple competition, and it is not the case that there is a unique system acting at a time. Rather, they are both active and take part in the decision proportionally to the fixed parameter ω. A simple switch between systems would not account for the full spectrum of observed behaviours ranging from STs to GTs [Mey+12].

Action selection

We use a softmax rule on the integrated values $\mathcal{P}$ to compute the probability to select an action A in state s:

$$p(a=A) = \frac{e^{\mathcal{P}(s,A)/\beta}}{\sum_{a'} e^{\mathcal{P}(s,a')/\beta}} \tag{5.8}$$

where $\beta > 0$ is the selection temperature that defines how probabilities are distributed. A high temperature ($\beta \rightarrow \infty$) makes all actions equiprobable, a low one makes the most rewarding action almost exclusive.

Impact of flupentixol

When simulating the pharmacological experiments, namely the impact of flupentixol, a parameter $0 \leq f < 1$ is used to represent the impact of flupentixol on parts of the model.

As a dopamine receptor antagonist, we model the impact of flupentixol on phasic dopamine by revising any RPE δ used in the model given the following formula:

$$\delta_f \leftarrow \begin{cases} \delta - f & \text{if } \frac{\delta - f}{\delta} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

where δ_f is the new RPE after flupentixol injection. The impact is filtered ($\frac{\delta-f}{\delta} \geq 0$) such that flupentixol injection could not lead to negative learning when the RPE was positive, but at most block it (i.e. the sign of δ_f cannot be different from the one of δ). With a low $f \rightarrow 0$, the RPE is not affected ($\delta_f \rightarrow \delta$). A high $f \rightarrow 1$ reduces the RPE, imitating a blockade of dopamine receptors.

Various studies (e.g. [Hum+12]) also suggest that tonic dopamine has an impact on action selection such that any decrease in dopamine level results in favouring exploration over exploitation. We therefore simulated the effect of flupentixol on action selection by revising the selection temperature given the following formula:

$$\beta_f \leftarrow \frac{\beta}{1-f} \tag{5.10}$$

where β_f is the new selection temperature, and $0 \leq f < 1$ represents the strength of the flupentixol impact. A strong $f \rightarrow 1$, which represents an effective dose of flupentixol, favours a high temperature $\beta_f \rightarrow \infty$ and therefore exploration. A low $f \rightarrow 0$, i.e. a low dose or an absence of flupentixol, leaves the temperature unaffected: $\beta_f \rightarrow \beta$.

For the first pharmacological experiment (Effects of systemic flupentixol administration on the learning of sign- and goal-tracking behaviours) both the impact on the softmax and on the RPE were activated, as the flupentixol was injected systemically and assumed to diffuse in the whole brain. For the second experiment (Effects of local flupentixol administration on the expression of sign- and goal-tracking behaviours) only the impact on the RPE was activated, as the flupentixol was injected locally in the core of the nucleus accumbens. We hypothesize that the Feature-Model-Free system relies in the core of the nucleus accumbens whereas the selection process (softmax) does not.

Initialization

In the original experiments [Fla+07; Fla+11b], prior to the autoshaping procedure, rats are familiarized with the Skinner box and the delivery of food into the magazine. While the MDP does not account for such pretraining, we can initialize the model with values ($\mathcal{Q}_i(s_1, goL)$, $\mathcal{Q}_i(s_1, goM)$ and $\mathcal{Q}_i(s_1, exp)$) that reflect it (see the estimation of the model parameters). These initial values can be seen as extra parameters common to the model and its variants.

Variants

Given the modular architecture of the model, we were able to test different combinations of RL systems. Their analysis underlined the key mechanisms required for reproducing each result (see Figures S5.1, S5.2, S5.4 and S5.5). Figure 5.11 (B, C and D) schematically represents the analysed variants.

Most of the results rely on the action taken by the agent at the lever appearance. The action taken results from the values $\mathcal{P}(s_1, goL)$, $\mathcal{P}(s_1, goM)$ and $\mathcal{P}(s_1, exp)$, the computation of which differs in each of the variants described below.

Variant 1 : Model-Free / Feature-Model-Free

Variant 1 was tested to assert the necessity of the Model-Based system as part of the model to reproduce the results. Thus in Variant 1, the Model-Based system is replaced by a classical Model-Free system, Advantage learning [Bai93; DB02], while the Feature-Model-Free system remains unchanged (see Figure 5.11 B).

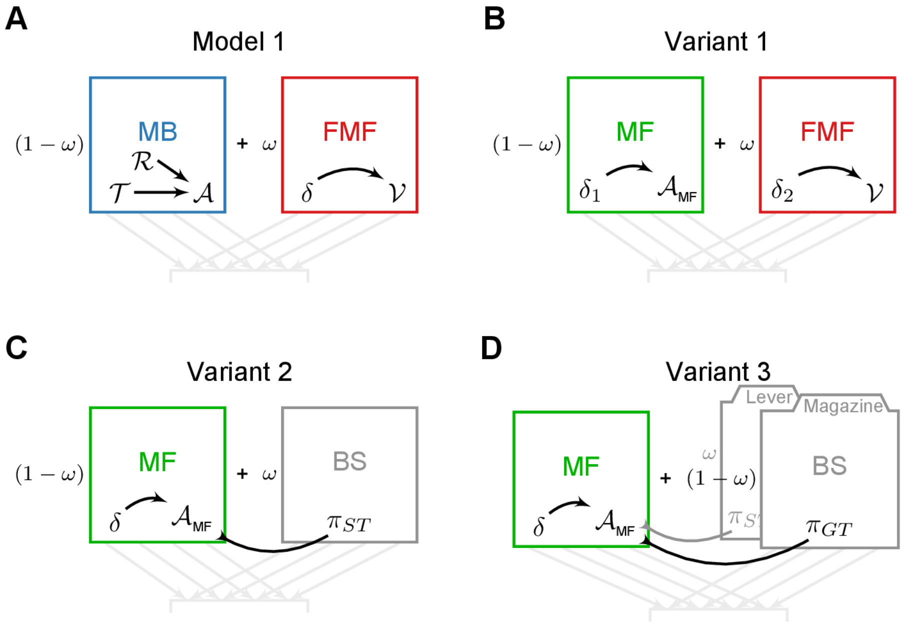


Figure 5.11: Systems combined in the model and the variants. *Variants of the model rely on the same architecture (described in Figure 5.2) and only differ in the combined systems. Colours are shared for similar systems.* **(A)** *The model combines a Model-Based system (MB, in blue) and a Feature-Model-Free (FMF, in red) system.* **(B)** *Variant 1 combines a Model-Free system (MF, in green) and a Feature-Model-Free system.* **(C)** *Variant 2 combines a Model-Free system and a Bias system (BS, in grey), that relies on values from the Model-Free system.* **(D)** *Variant 3 combines a Model-Free system and two Bias systems, that rely on values from the Model-Free system. Variant 4 is not included as it failed to even reproduce the autoshaping behavioural results.*

In such a Model-Free system, the action-value function $\mathcal{Q}_{\text{MF}}$ is updated online according to the transition just experienced. At each time step the function is updated given an RPE δ that computes the difference between the observed and the expected value, as follows:

$$\begin{aligned} \mathcal{Q}_{\text{MF}}(s,a) &\leftarrow \mathcal{Q}_{\text{MF}}(s,a) + \alpha\delta \\ \delta &\leftarrow r + \gamma \max_{a'} \mathcal{Q}_{\text{MF}}(s',a') - \mathcal{Q}_{\text{MF}}(s,a) \end{aligned} \tag{5.11}$$

Computation of the associated Advantage function $\mathcal{A}_{\text{MF}}$ follows Equation (5.5). This model computes integrated values as follows:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{\text{MF}}(s,a) + \omega\mathcal{V}(c(s,a)) \tag{5.12}$$

It is important to note that while Equation (5.12) looks similar to Equation (6.4), the Advantage function is computed by a Model-Based system in the model ($\mathcal{A}$) and a Model-

Free system in this variant ($\mathcal{A}_{\text{MF}}$), leading to very different results on pharmacological experiments.

Variant 2 : Asymmetrical

Inspired by a work from Dayan et al. [Day+06], Variant 2 combines a classical Advantage learning system [Bai93; DB02] with some Bias system taking its values directly from the other system (see Figure 5.11 C). This system computes the integrated values as follows:

$$\mathcal{P}(s,a) = (1-\omega) \times \mathcal{A}_{\text{MF}}(s,a) + \omega \begin{cases} \mathcal{V}(s) & \text{if } a = goL \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

It asymmetrically gives a bonus to the path that should be taken by STs. In slight discrepancy with the original model, it uses the maximum value over action-value function $\mathcal{Q}_{\text{MF}}$ as the value function $\mathcal{V}_{\text{MF}}$ used to compute the advantage function. Hence, there is a single RPE computed at each step.

Variant 3 : Symmetrical

In the same line as Variant 2, Variant 3 symmetrically gives a bonus to both paths using a classical Advantage learning system in combination with a Pavlovian system. This system computes the integrated values as follows:

$$\mathcal{P}(s,a) = \mathcal{A}_{\text{MF}}(s,a) + \begin{cases} \omega\mathcal{V}(s) & \text{if } a = goL \\ (1-\omega)\mathcal{V}(s) & \text{if } a = goM \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

This model does not exactly fit Equation (6.4) of the general architecture. It is based on 3 systems, where the real competition is between the two bias systems, whereas the Model-Free system is mainly used to compute the values used by the two others (see Figure 5.11 D). The rest of the architecture is not impacted.

Variant 4 : Model-Based / Model-Free

Variant 4 was developed to confirm the necessity of a feature-based system. It combines two advantage functions computed from a Model-Based ($\mathcal{A}$) and a Model-Free ($\mathcal{A}_{\text{MF}}$) system.

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}(s,a) + \omega\mathcal{A}_{\text{MF}}(s,a) \tag{5.15}$$

While computed differently, both advantage functions will eventually converge to the same optimal values [SB98] making both systems favouring the same optimal policy. Note that u_{ITI} cannot be used in this variant as there exists no value over the magazine itself. While varying the parameters might slow down learning or make the process more exploratory, this could never lead to sign-tracking as both systems, whatever the weighting, would favour goal-tracking. As such, Variant 4 is unable to even account for the main behavioural results in the autoshaping procedure (see Figure S5.8).

Given that all the subsequent simulated results relies on a correct reproduction of the default behaviours, this variant was not investigated further and is not compared to the other variants in supplementary results figures.

Estimating the model parameters

The model relies on model-specific parameters (ω, β, α and γ) and experience-specific parameters (u_{ITI}, $\mathcal{Q}_i(s_1, \text{goL})$, $\mathcal{Q}_i(s_1, \text{goM})$ and $\mathcal{Q}_i(s_1, \emptyset)$). If the model were used to simulate a different experiment, the model-specific parameters would be the same while different experience-specific parameters might be required. For an easier analysis and a simpler comparison between the model and its variants, we reduce the number of parameters by sharing parameters with identical meanings amongst systems (i.e. both systems within the model share values for their learning rates α and discount rates γ, rather than having independent parameter values).

Due to the number of parameters, finding the best values to qualitatively fit the experimental data cannot be done by hand. Using a genetic algorithm makes it possible to optimize the search of suitable values for the parameters.

Parameter values were retrieved by fitting the simulation of the probabilities to engage either the lever or the magazine with the experimental data of one of the previous studies [Fla+09]. No direct fitting was intended on other experimental data. Hence, a single set of values was used to simulate behavioural, physiological and pharmacological data.

If for a variant, the optimization algorithm fails to fit the experimental data, it suggests that whatever the values, the mechanisms involved cannot explain the behavioural data (Variant 4).

Probabilities to engage the lever or the magazine were taken as independent objectives of the algorithm, since fitting sign-tracking probabilities is easier than fitting goal-tracking probabilities. For each objective, the fitness function is computed as the least square errors between the experimental and simulated data. Parameter optimization is done with the multi-objective genetic algorithm NSGA-II [Deb+02]. We used the implementation provided by the Sferes 2 framework [MD10]. All parameters required for reproducing the behavioural data were fitted at once.

For NSGA-II, we arbitrarily use a population of 200 individuals and run it over 1000 generations. We use a polynomial mutation with a rate of 0.1, and simulate binary crossovers with a rate of 0.5. We select the representative individual, to be displayed in figures, from the resulting Pareto front by hand, such that it best visually fits the observed data.

To confirm that ω is the key parameter of the model, we additionally tried to fit the whole population at once (i.e. sharing all parameter values in agents but ω) and we were still able to reproduce the observed tendencies of sign- and goal-tracking in the population (see Figure S5.7 A,B) and the resulting different phasic dopaminergic patterns (see Figure S5.7 C,D).

It is however almost certain that each subgroup does not express the exact same values for the other parameters. Removing such constraint by fitting each subgroup separately, indeed provides better results. Results presented in this article are based on such separate fitting.

Supporting Information

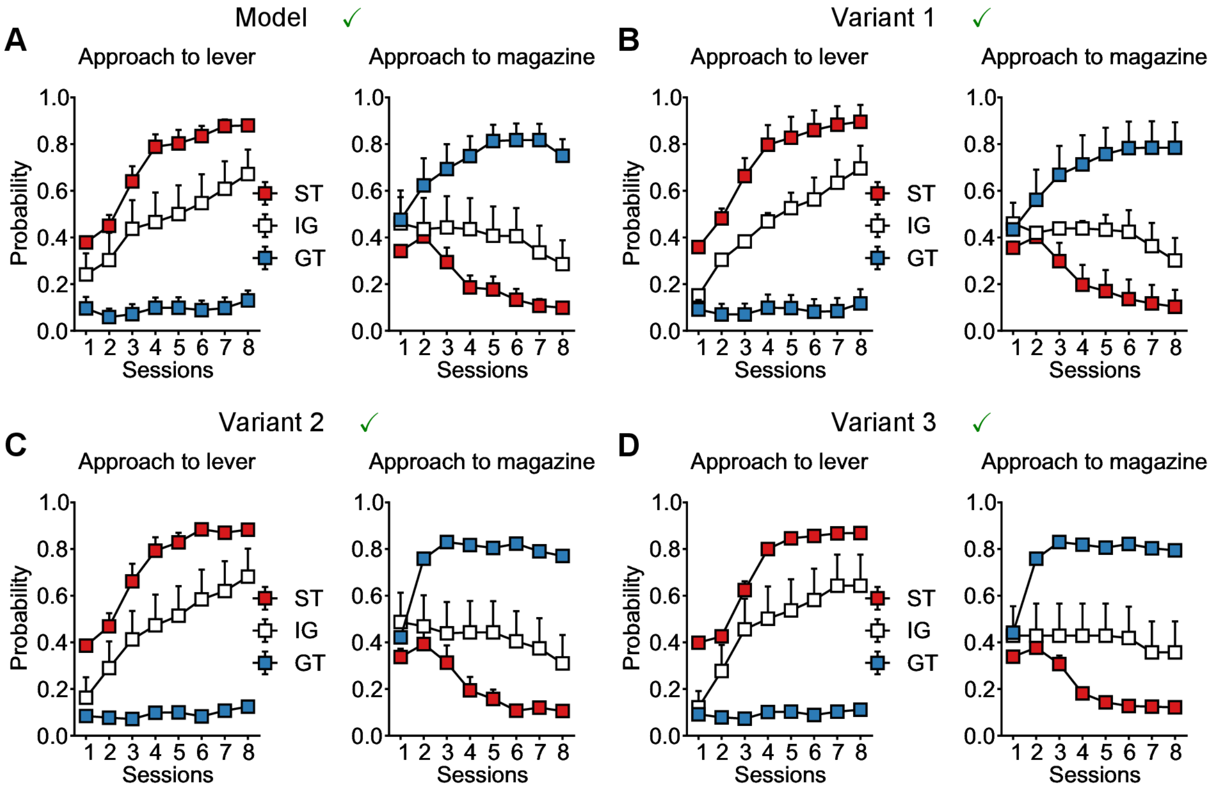


Figure S5.1: Comparison of variants of the model on simulations of autoshaping experiment. *Legend is as in Figure 5.5 (C,D). Simulation parameters for STs (red), GTs (blue) and IGs (white) in the model* **(A)***, Variant 1* **(B)***, Variant 2* **(C)** *and Variant 3* **(D)** *are summarized in Table S5.1. All variants reproduce the spectrum of behaviours ranging from sign-tracking to goal-tracking.*

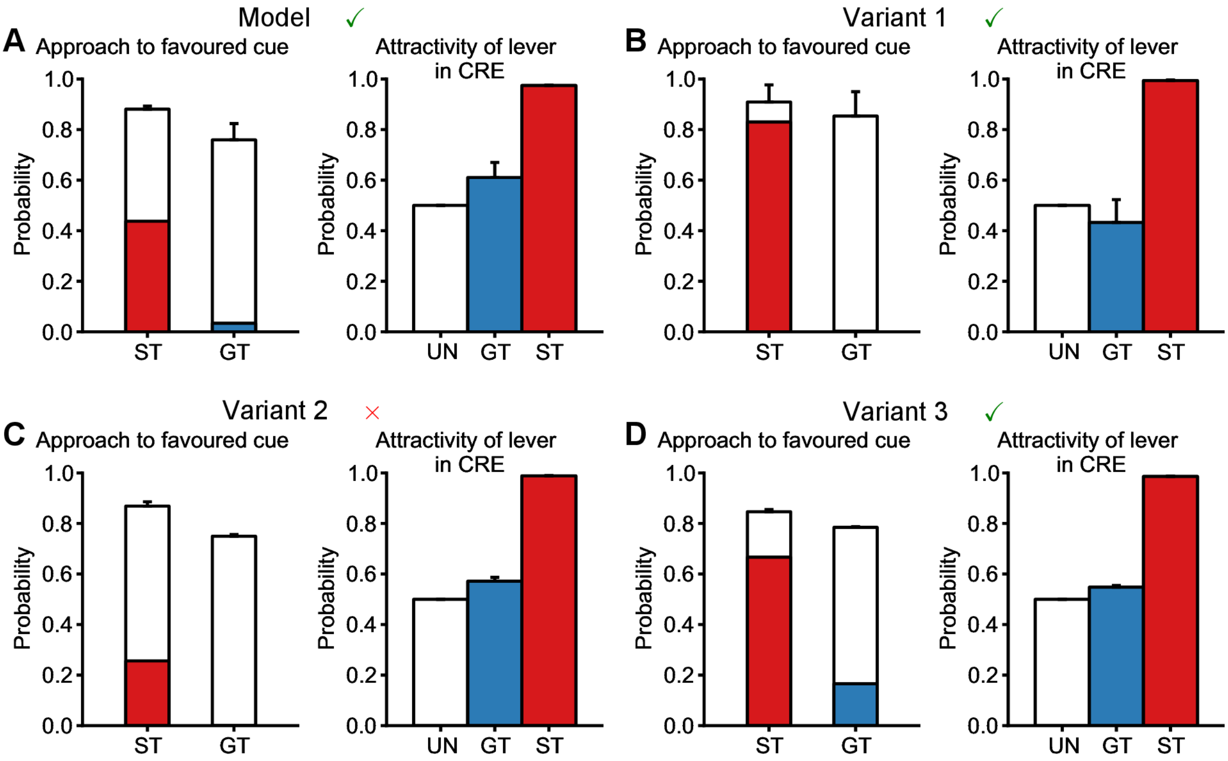


Figure S5.2: Comparison of variants of the model on incentive salience and Conditioned Reinforcement Effect intuitions. *Legend is as in Figure 5.6. Simulation parameters for STs (red), GTs (blue) and IGs (white) are summarized in Table S5.1. Variant 2* **(C)** *relying on asymmetrical bonuses given only to sign-tracking cannot reproduce the attribution of a motivational value by the second system to both the lever and the magazine. Others* **(A,B,D)** *attribute values to both stimuli and parallels the supposed acquisition of motivational values by stimuli, i.e. incentive salience. All variants are able to account for a Conditioned Reinforcement Effect more pronounced in STs than in GTs.*

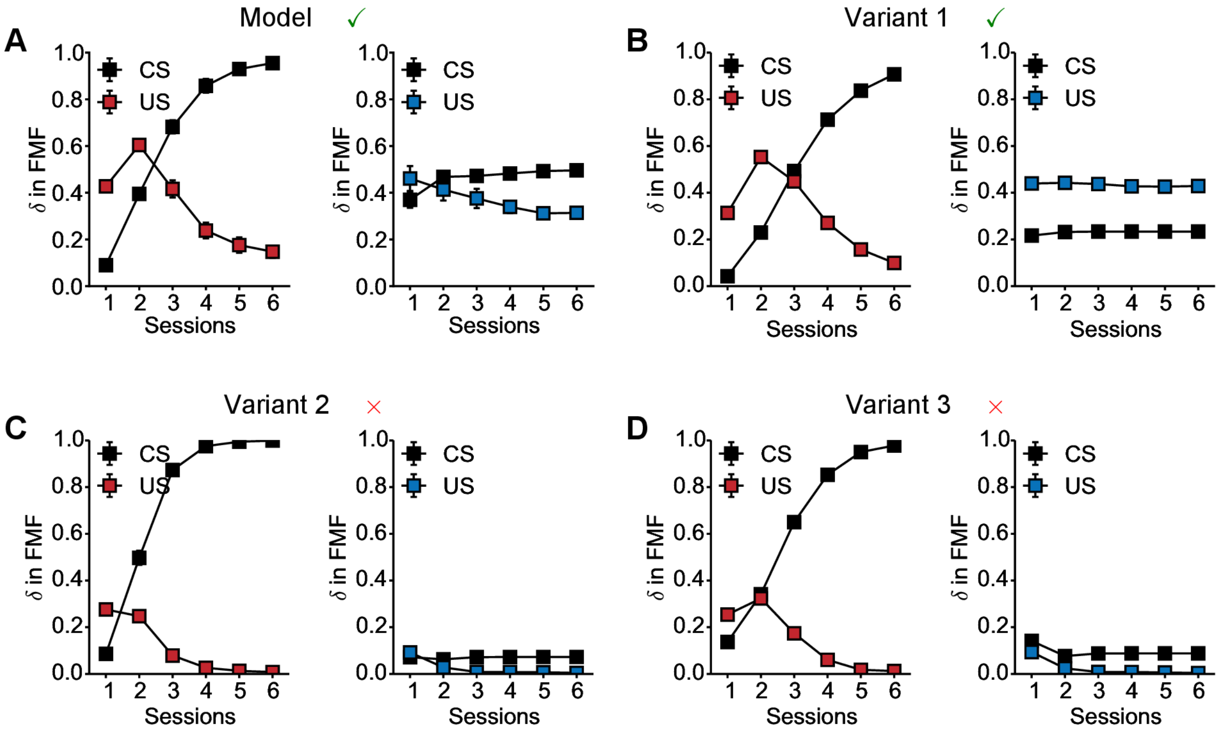


Figure S5.3: Comparison of variants of the model on simulations of patterns of dopaminergic activity. *Legend is as in Figure 5.7 (C,D). Simulation parameters for STs (left) and GTs (right) are summarized in Table S5.1. The model* **(A)** *and Variant 1* **(B)** *can reproduce the difference observed in dopaminergic patterns of activity in STs versus GTs. Other variants* **(C,D)** *fail to do so, given that the classical Model-Free system propagates the RPE from food delivery to lever appearance on all pathways of the MDP.*

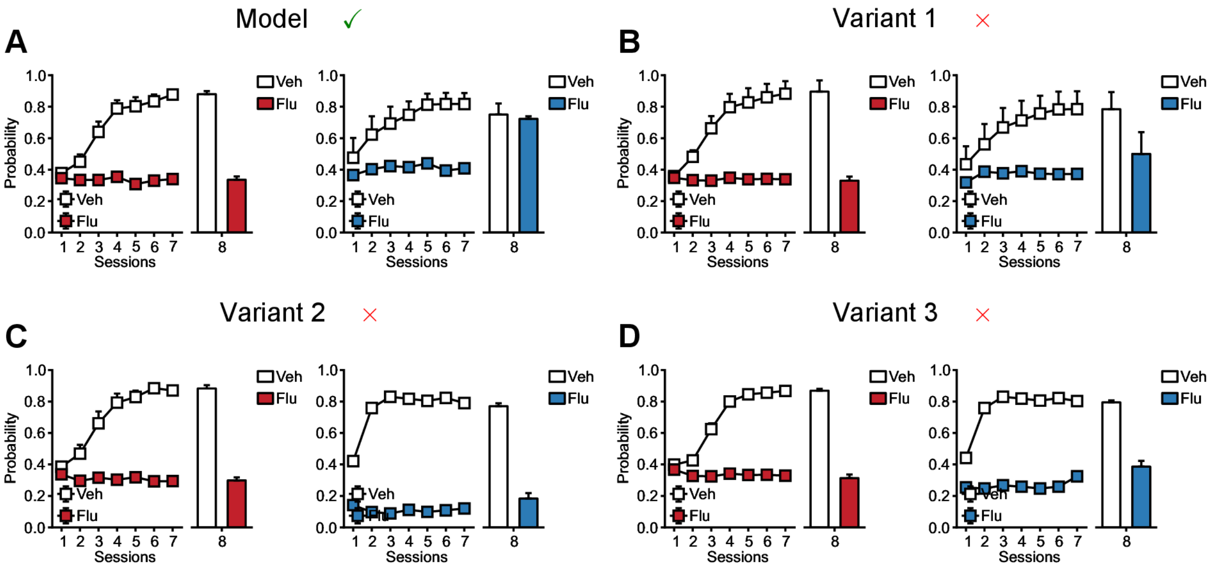


Figure S5.4: Comparison of variants on simulations of the effect of systemic injections of flupentixol. *Legend is as in Figure 5.8 (C,D). Simulation parameters for STs (left) and GTs (right) are summarized in Table S5.1. Only the Model* **(A)** *can reproduce the difference in response to injections of flupentixol observed in STs versus GTs. All variants* **(B,C,D)** *fail to do so, given that they only rely on Model-Free, i.e. RPE-dependent, mechanisms that are blocked by flupentixol.*

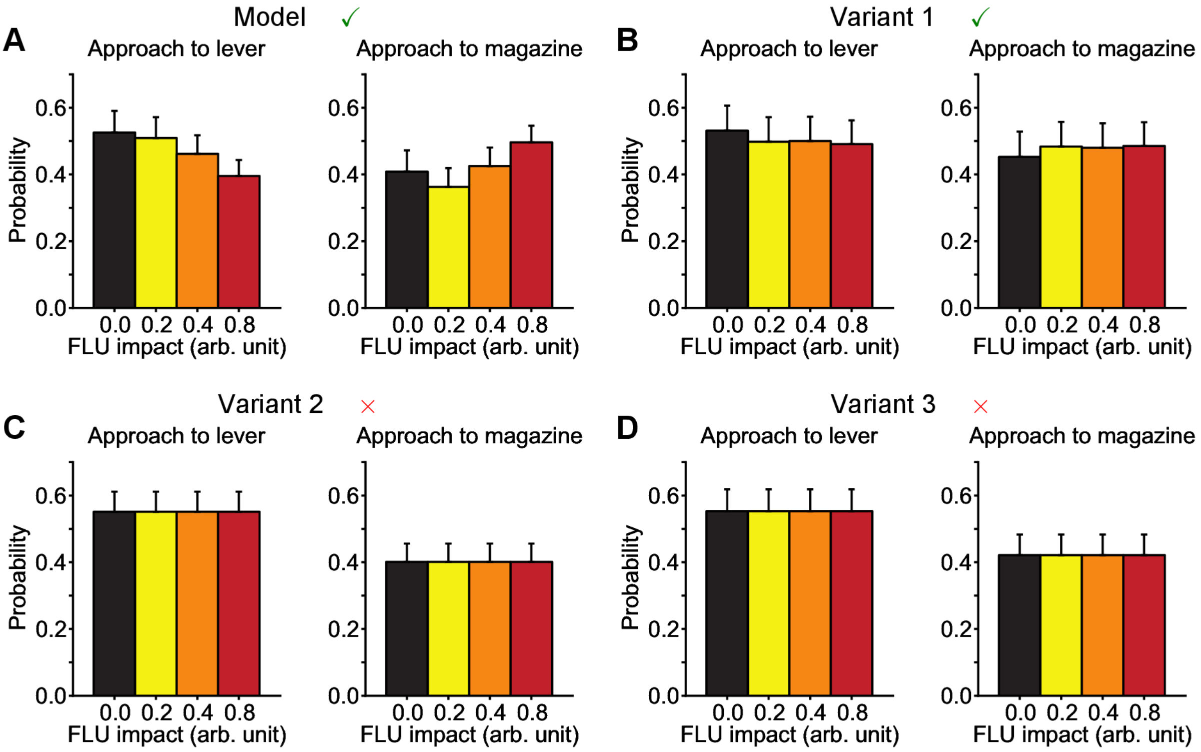


Figure S5.5: Comparison of variants on simulations of the effect of post injections of flupentixol. *Legend is as in Figure 5.9 (C,D). Simulation parameters for groups of rats composing the population are summarized in Table S5.1. Variants 2* **(C)** *and 3* **(D)***, accounting for sign- and goal-tracking using a single set of values, have a similar impact of flupentixol on both behaviours, leaving relative probabilities to engage with lever and magazine unaffected. Variant 1* **(B)** *uses different systems, thus flupentixol impacts sign-tracking in the model in the same way as it does in experimental data. However, given that both systems rely on RPE-dependent mechanisms, the impact is not as visible as in the model* **(A)***.*

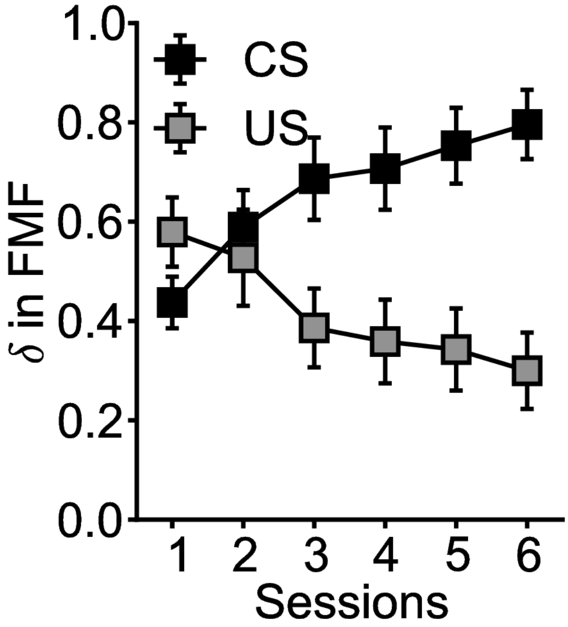


Figure S5.6: Prediction of the model about expected patterns of dopaminergic activity in intermediate groups. *Data are expressed as mean ± S.E.M. Average RPE computed by the Feature-Model-Free system in response to CS and US presentation for each session of conditioning in the intermediate group. Simulated group is defined as in Figure 5.5.*

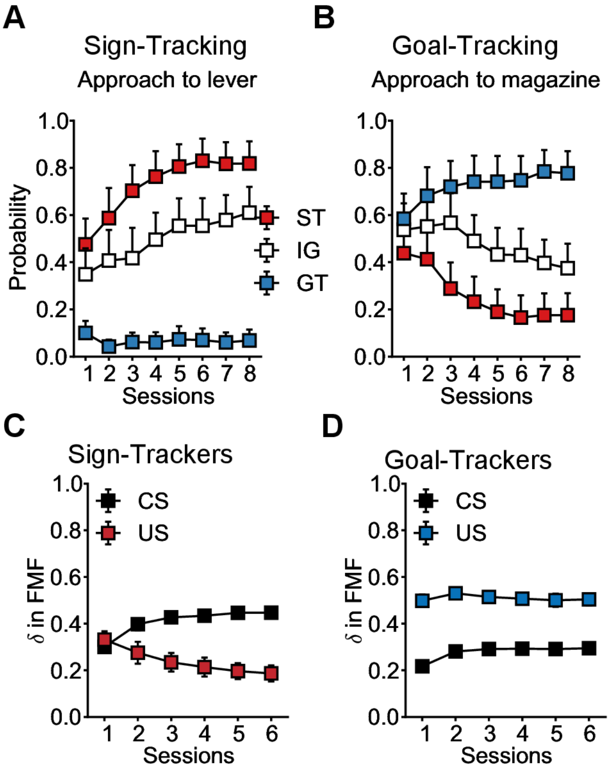


Figure S5.7: Behavioural and Physiological simulations of autoshaping with shared parameter values across STs, GTs and IGs. **(A,B)** *Legend is as in Figure 5.5 (C,D). Reproduction of the respective tendencies to sign- and goal-track of STs ($\omega = 0.5$), IGs ($\omega = 0.375$) and GTs ($\omega = 0.05$)) using a single set of parameters ($\alpha = 0.2$, $\gamma = 0.8$, $\beta = 0.09$, $u_{ITI} = 0.2$, $\mathcal{Q}_i(s_1, goL) = 0.0$, $\mathcal{Q}_i(s_1, exp) = 0.5$ and $\mathcal{Q}_i(s_1, goM) = 0.5$).* **(C,D)** *Legend is as in Figure 5.7 (C,D). Reproduction of the different patterns of phasic dopaminergic activity in STs and GTs using the same single set of parameters. By simply varying the ω parameter, the model can still qualitatively reproduce the observations in experimental data.*

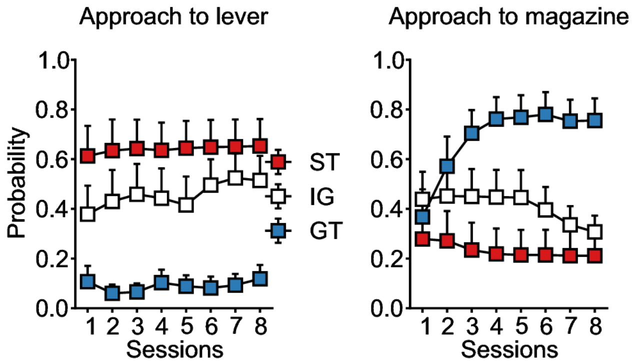


Figure S5.8: Simulation of autoshaping experiment for Variant 4. *Legend is as in Figure 5.5 (C,D). Simulation for parameters STs (red), GTs (blue) and IGs (white) in the Variant 4 are summarized in Table S5.1. Variant 4 is not even able to reproduce the main behavioural data.*

Table S5.1: Summary of parameters used in simulations

Version	Type	ω	β	α	γ	u_{ITI}	$\mathcal{Q}_i(s_1, L)$	$\mathcal{Q}_i(s_1, \emptyset)$	$\mathcal{Q}_i(s_1, M)$
A: Model	ST	0.499	0.239	0.031	0.996	0.027	0.844	0.999	0.538
	IG	0.276	0.142	0.217	0.999	0.228	0.526	0.888	0.587
	GT	0.048	0.084	0.895	0.727	0.140	1.0	0.316	0.023
B: Variant 1	ST	0.994	0.145	0.018	0.999	0.995	0.278	0.999	0.676
	IG	0.350	0.095	0.023	0.971	0.904	0.398	0.675	0.712
	GT	0.003	0.002	0.906	0.508	0.263	0.147	0.419	0.520
C: Variant 2	ST	0.788	0.367	0.055	0.996	0	0.153	0.133	0.151
	IG	0.843	0.046	0.779	0.999	0	0	0.532	0.593
	GT	0.211	0.130	0.109	0.445	0	0	1	0.095
D: Variant 3	ST	0.295	0.189	0.070	0.999	0	0.057	0.054	0
	IG	0.333	0.027	0.926	0.674	0	0.011	0.444	0.747
	GT	0.166	0.047	0.093	0.417	0	0	0.476	0.229
E: Variant 4	ST	0.643	0.136	0.763	1	-	0.325	0.713	0.094
	IG	0.277	0.175	0.748	0.999	-	0.273	0.784	0.986
	GT	0.529	0.077	0.617	0.695	-	0.102	0.635	0.962
F: Model (shared)	ST	0.500	0.090	0.20	0.800	0.200	0.000	0.400	0.400
	IG	0.375	0.090	0.20	0.800	0.200	0.000	0.400	0.400
	GT	0.050	0.090	0.20	0.800	0.200	0.000	0.400	0.400

Parameters retrieved by optimisation with NSGA-II and used to produce the results presented in this article for the model and its variants. Parameters for STs, GTs and IGs were optimized separately (A,B,C,D,E). To confirm that ω is the key parameter of the model, we also optimized parameters for STs, GTs and IGs by sharing all but the ω parameter (F) to produce Figure S5.7.

Table S5.2: Definition of feature-function c

s	s_0	s_1	s_1	s_1	s_2	s_3	s_4	s_5	s_6	s_7
a	exp	goL	exp	goM	eng	$\emptyset$	eng	goM	goM	eat
$c(s,a)$	$\emptyset$	L	$\emptyset$	M	L	$\emptyset$	M	F	F	F

Stimuli (*L*ever, *M*agazine, *F*ood or $\emptyset$) returned by the feature-function c for each possible state-action pair $\langle s, a\rangle$ in the MDP described in Figure 5.1. The feature-function simply defines the stimulus that is the focus of an action in a particular state.

Chapter 6

This chapter lists some predictions drawn from the computational model developed in Chapter 5. This work is presented under the form of a published journal paper:

> *Florian Lesaint, Olivier Sigaud, Jeremy J Clark, Shelly B Flagel, and Mehdi Khamassi. "Experimental predictions drawn from a computational model of sign-trackers and goal-trackers". In:* J Physiol Paris *(2014). in press*
>
> http://www.sciencedirect.com/science/article/pii/S0928425714000242

We develop in details predictions scarcely evoked in the discussion of our first paper [Les+14a]. We suggest new data analyses and revised experimental setups. If conducted, such studies would have the power to reinforce or refute the model, and would definitely contribute to the understanding of the biological mechanisms involved in the present task. It aims at comforting the model, especially regarding the key concepts it combines, that is dual-learning systems and factored representations, and regarding hypotheses on the impact of the inter-trial interval.

The predictions suggest that dopaminergic patterns for IGs should be a mixed signal between those observed for STs and GTs. We predict that looking separately at the DA patterns given the prior engagement towards either the lever or the magazine should lead to clearly distinct patterns. We predict that the removal of the magazine during the ITI should lead to an increased motivational engagement towards the magazine, a decreased tendency in sign-tracking within the population and a different pattern of dopaminergic activity when goal-tracking. Finally, we predict that local injections of flupentixol to the core of the nucleus accumbens would preserve goal-tracking and prevent the learning of a sign-tracking response, a result that should also be observed following lesions of the core of the nucleus accumbens prior to conditioning. Lesions after conditioning would only block the expression of the learned sign-tracking behaviour.

Experimental predictions drawn from a computational model of sign-trackers and goal-trackers

Florian Lesaint^{1,2,*}, Olivier Sigaud^{1,2}, Jeremy J. Clark^{3}, Shelly B. Flagel$^{4-6}$, Mehdi Khamassi^{1,2}

1 Sorbonne Universités, UPMC Univ Paris 06, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France

2 CNRS, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France

3 Department of Psychiatry and Behavioral Sciences, University of Washington, Washington, USA

4 Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, United States of America

4 Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, Michigan, United States of America

5 Department of Psychology, University of Michigan, Ann Arbor, Michigan, United States of America

$\ast$ E-mail: Corresponding lesaint@isir.upmc.fr

Abstract

Gaining a better understanding of the biological mechanisms underlying the individual variation observed in response to rewards and reward cues could help to identify and treat individuals more prone to disorders of impulsive control, such as addiction. Variation in response to reward cues is captured in rats undergoing autoshaping experiments where the appearance of a lever precedes food delivery. Although no response is required for food to be delivered, some rats (goal-trackers) learn to approach and avidly engage the magazine until food delivery, whereas other rats (sign-trackers) come to approach and engage avidly the lever. The impulsive and often maladaptive characteristics of the latter response are reminiscent of addictive behaviour in humans. In a previous article, we developed a computational model accounting for a set of experimental data regarding sign-trackers and goal-trackers. Here we show new simulations of the model to draw experimental predictions that could help further validate or refute the model. In particular, we apply the model to new experimental protocols such as injecting flupentixol locally into the core of the nucleus accumbens rather than systemically, and lesioning of the core of the nucleus accumbens before or after conditioning. In addition, we discuss the possibility of removing the food magazine during the inter-trial interval. The predictions from this revised model will help us better understand the role of different brain regions in the behaviours expressed by sign-trackers and goal-trackers.

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

Introduction

A significant number of models have been developed since the 1970s to describe Pavlovian and instrumental phenomena. Early models were mostly focusing on reproducing the averaged behaviour expressed within a population, neglecting inter-individual variations and possibly smoothing the true behaviour of individuals [Gal+04], or even masking the variation in behaviour. However, this variation is of particular interest when trying to identify those individuals within population prone to impulsive behaviours or having a higher risk of addiction [Fla+11b; SR13; Huy+14].

Recent studies have investigated such intervariability among rats undergoing an autoshaping experiment [Fla+07; Fla+09; Fla+11b; Fla+11a; DB12b; MB09; RF09; Mey+12; Fit+13], where a lever (conditioned stimulus, CS) was presented for 8 seconds, followed immediately by delivery of a food pellet (unconditioned stimulus, US) into an adjacent food magazine. Although no response was required to receive the reward, with training, some rats (sign-trackers; STs) learned to rapidly approach and engage the lever-CS. However, others (goal-trackers; GTs) learned to approach the food magazine upon CS presentation, and made anticipatory head entries into it. Some rats (intermediate group; IG) presented a mixed behaviour, switching between lever and magazine during presentation of the CS, and sometimes engaging both during one trial. Furthermore, in STs, phasic dopamine release in the core of the nucleus accumbens, measured with Fast Scan Cyclic Voltammetry (FSCV), matched the pattern that would be predicted by reward prediction error (RPE) signalling, and dopamine was necessary for the acquisition of a sign-tracking conditioned response (CR). In contrast, despite the fact that GTs acquired a Pavlovian conditioned approach response, this was not accompanied with the expected RPE-like dopamine signal, nor was the acquisition of a goal-tracking CR blocked by administration of a dopamine antagonist (see also Danna and Elmer [DE10]). While the proportion of STs and GTs in the population varies [Fit+13], both phenotypes are typically represented in an outbred population.

To our knowledge, only one model [Les+14b] accounts for these experimental results and has been validated with existing data. This model is built on a combination of model-free and model-based systems [Daw+05; Cla+12; Huy+14] and extended with state factored representations. Combining multiple systems enables the model to express a large repertoire of behaviours and considering features within states enables the model to learn Pavlovian impetuses [Day+06] specific to the Pavlovian features within the task.

In this paper, we review the model described by Lesaint et al. [Les+14b], extending it with a new tool to improve its reliability. We suggest new experimental protocols and some new analyses of the data that would further validate the model and strengthen its explanatory power, refine our understanding of the role of the nucleus accumbens in the described behaviours, and help clarify the impact of some choices made in the original protocol.

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

Material and methods

The model from which the present results are extracted is described in depth in a previous article [Les+14b]. It is composed of two distinct reinforcement learning systems that collaborate to define the action to be selected at each step of the experiment (see Figure 6.1 A; Clark et al. [Cla+12]).

The first system, a model-based system (MB), incrementally learns a model of the world (a transition function $\mathcal{T}$ and a reward function $\mathcal{R}$) from which it infers values ($\mathcal{A}$) for each action in each situation, given the classical following formulas:

$$\mathcal{Q}(s,a) \leftarrow \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \max_{a'} \mathcal{Q}(s',a') \tag{6.1}$$

$$\mathcal{A}(s,a) \leftarrow \mathcal{Q}(s,a) - \max_{a'} \mathcal{Q}(s,a') \tag{6.2}$$

where the discount rate $0 \leq \gamma \leq 1$ classically represents the preference for immediate versus distant rewards. At each step, the most valued action is the most rewarding on the long run (e.g. approaching the magazine to be ready to consume the food as soon as its delivery). It favours goal-tracking because this is the shortest path towards the rewarding state (see Figure 6.1 B).

The second system, a revised model-free system, learns values ($\mathcal{V}$) over features (e.g. food, lever or magazine). Contrary to the first system, which uses a classical abstract state representation, it relies on the features that compose these abstract states. In traditional reinforcement learning, each situation that can be encountered by the agent is defined as an abstract state (e.g. arbitrarily defined as s_1, $s_2 \ldots s_x$), such that similarities between situations (e.g. presence of a magazine) are lost. By using features, we reintroduce the capacity to use and benefit from these similarities. The second system is further defined as the feature model-free system (FMF). It relies on a RPE signal δ, computed as follows:

$$\begin{aligned} \mathcal{V}(c(s,a)) &\leftarrow \mathcal{V}(c(s,a)) + \alpha\delta \\ \delta &\leftarrow r + \gamma \max_{a'} \mathcal{V}(c(s',a')) - \mathcal{V}(c(s,a)) \end{aligned} \tag{6.3}$$

where $c : \mathcal{S} \times \mathcal{A} \rightarrow \{lever, magazine, food, \emptyset\}$ is a feature-function that returns the feature $c(s,a)$ the action a was focusing on in state s (e.g. it returns the lever when the action was to engage with the lever). We hypothesized that, similarly to classical model-free systems, δ parallels phasic dopaminergic activity [Sch98]. This signal enables to revise and attribute values, seen as motivational, to features without the need of the internal model of the world used by the MB system. When an event is fully expected, there should be no RPE as its value is fully anticipated. When an event is positively surprising, there should be a positive RPE. Actions are then valued by the motivational value of the feature they are focusing on (e.g. engaging with the lever would be valued given the general motivational value of the lever). Hence, it favours actions that engage with the most motivational features. This might lead to favour suboptimal actions with regard to maximizing rewards (e.g. engaging with the lever keeps the rat away from the soon to be rewarded magazine). It favours sign-tracking (a suboptimal path, see Figure 6.1 B) as the lever, being a full predictor of reward, earns a strong motivational value relative to the magazine.

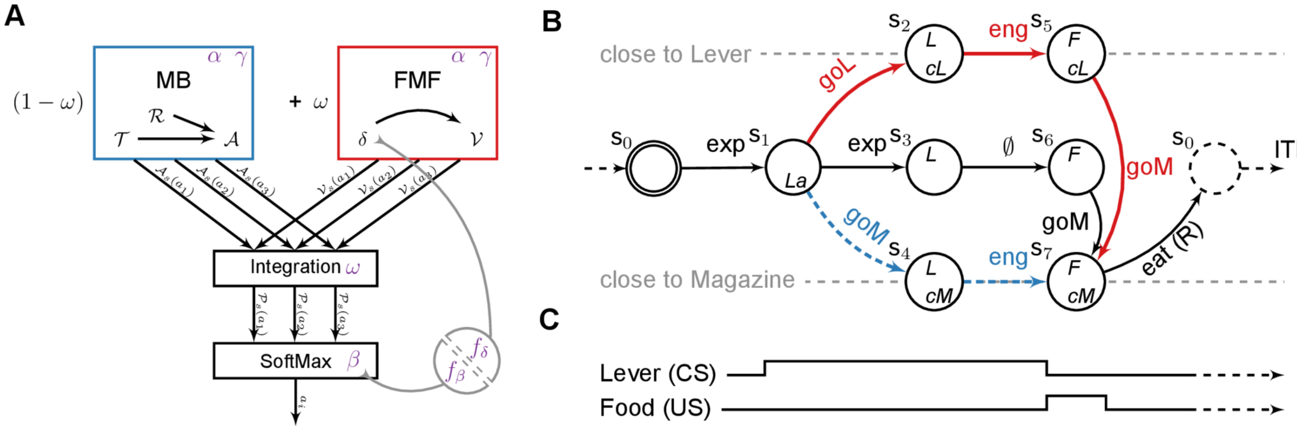


Figure 6.1: Model and Markov Decision Process used for simulations. **(A)** *The model is composed of a model-based system (MB, in blue) and a Feature-Model-Free system (FMF, in red) which provide respectively an advantage function $\mathcal{A}$ and a value function $\mathcal{V}$ values for actions a_i given a state s. These values are integrated in $\mathcal{P}$, prior to be used into an action selection mechanism. The various elements may rely on parameters (in purple). The impact of flupentixol on dopamine is represented by a parameter f that influences the action selection mechanism and/or any reward prediction error that might be computed in the model.* **(B)** *MDP accounting for the experiments described in Flagel et al. [Fla+09; Fla+11b]; Robinson and Flagel [RF09]; Meyer et al. [Mey+12]. States are described by a set of variables: L/F - Lever/Food is available, cM/cL - close to the Magazine/Lever, La - Lever appearance. The initial state is double circled, the dashed state is terminal and ends the current episode. Actions are* engage *with the proximal stimuli,* explore*, or* go *to the* Magazine*/*Lever *and* eat*. The path that STs should favour is in red. The path that GTs should favour is in dashed blue.* **(C)** *Time line corresponding to the unfolding of the MDP.*

The model does not base its decision on a single system at a time, rather the values of the MB system ($\mathcal{A}_{MB}$) and the FMF system ($\mathcal{V}_{FMF}$) are integrated such that a single decision is made at each time step: producing a sort of cooperation between the two systems. The values computed by these two systems are then integrated through a weighted sum and passed to a softmax action selection mechanism that converts them into probabilities of selecting the action given a situation (see Figure 6.1 A). The integration is done as follows:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{MB}(s,a) + \omega\mathcal{V}_{FMF}(c(s,a)) \tag{6.4}$$

where $0 \leq \omega \leq 1$ is a combination parameter which defines the importance of each system in the overall model. Varying ω (while leaving the other parameters of the model unchanged) is sufficient to reproduce the characteristics of the different subgroups of rats [Les+14b]. The previous experimental data could be reproduced by having STs give a

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

stronger weight to the FMF system whereas having GTs give a stronger weight to the MB system. FMF and MB systems are then updated according to the action a taken by the full model in state s - even if the systems would have individually favoured different actions -and the resulting new state s' and retrieved reward r, as previously done in other computational models involving a cooperation between model-free and model-based systems [Cal+12].

Simulations of experimental protocols

The experiment is described through an episodic Markov Decision Process (MDP) that represents one trial of the session (see Figure 6.1 B,C). The inter-trial interval (ITI), not being part of the MDP, is simulated between each run by revising downward the magazine value ($\mathcal{V}_{FMF}(M) \leftarrow (1 - u_{ITI}) \times \mathcal{V}_{FMF}(M)$, u_{ITI} being a parameter of the model). This simulates the hypothesis that the presence of the magazine in the absence of food delivery reduces its value. If the magazine were removed during ITI, we would expect no revision of its value.

The model is used to simulate experiments that involved injections of flupentixol, an antagonist of dopamine, either systemically or within the core of the nucleus accumbens. In the case of local injections, assuming that the FMF system relies on the core of the nucleus accumbens, we simulate the impact of flupentixol on phasic dopamine by degrading the reward predictions errors as follows:

$$\delta \leftarrow \begin{cases} \delta - f & \text{if } \frac{\delta - f}{\delta} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

where $0 \leq f < 1$ represents the impact of flupentixol. Its effect is defined such that flupentixol injections cannot lead to negative learning when RPE is positive, but at most blocks it. In the case of systemic injections, we also assume an additional impact on tonic dopamine [Hum+12], which affects the action selection process. We simulate this impact by revising the temperature parameter ($\beta \leftarrow \frac{\beta}{1-f}$). Hence, flupentixol favours random exploration instead of using learned values to take a decision.

Some predictions presented here suggest to lesion the core of the nucleus accumbens. Such a lesion is simulated by removing the FMF system from the model , i.e. all values that would have come from the system are replaced by 0. The rest of the model is left intact. Equation 6.4 can be replaced by:

$$\mathcal{P}(s, a) = (1 - \omega)\mathcal{A}_{MB}(s, a) \tag{6.6}$$

Index Score

Introduced by Meyer et al. [Mey+12], the Pavlovian Conditioned Approach (PCA) Index Score provides a metric to categorize rats as STs, GTs or IGs independent of the rest of the population. That is, instead of ordering rats based on their engagement with the lever and splitting the population in 3 groups of approximately equal size, as done in previous studies [Fla+07; RF09], classifying rats based on PCA Index minimizes the

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

chances of misclassification and allows one to compare across studies or populations of rats. The PCA Index relies on the number of contacts with the lever and the magazine, the probability to engage with one versus the other and the latencies to act towards each (Table 1 in [Mey+12]).

We developed a similar Index Score as it provides a good metric for some of the predictions described here. Simulated rats whose score is > 0.5 are defined as STs. Simulated rats that have a score < -0.5 are defined as GTs. Remaining rats are defined as IGs. Table 6.1 explains how it is computed based on the last two sessions of simulations. Contrary to the PCA Index Score, it cannot use latencies as they are not accounted for by the model.

Response Bias(n)	$= (LeverPresses – MagazineEntries)/(LeverPresses + MagazineEntries)$
Probability Difference(n)	$= p(LeverPress) – p(MagazineEntry)$
Score(n)	$= [ResponseBias(n) + ProbabilityDifference(n)]/2$
Index Score	$= [Score(6) + Score(7)]/2$

Table 6.1: Formulas for deriving the Index Score. *The Index Score provides a way to classify rats as STs, GTs or IGs, independently of the rest of the population. It relies on averaging scores computed for the last two sessions of the simulations. The Score for session n is derived by averaging its Response Bias and its Probability Difference. Responses Bias is a ratio between the difference in lever presses versus magazine entries and the total number of entries. Probability Difference is the difference between the probability to engage with the lever and the probability to engage with the magazine.*

Estimation of model parameters

The model relies on a set of 8 parameters (a shared learning rate, a shared discount rate, a selection temperature, an integration parameter, a ITI impact parameter and 3 initial conditions) that need to be tuned for simulations to fit experimental data. We use the multi-objective algorithm NSGA-II [Deb+02; MD10] to find the best values (solutions) for the parameters. This method is an efficient tool to fully explore the high dimensional parameter space and avoid local minima.

As in [Les+14b], we search a set of parameter values per group. The two first objectives of the fitness function are to fit the averaged behaviours of the simulated group to the averaged behaviours of the experimental group. More formally, for each group, we try to minimize the least square error between the probabilities of rats and simulated rats to engage with the magazine and the lever over time (see Table 6.2). This results in multiple solutions that are compromises between these two objectives. We subsequently select one of the solutions that is visually acceptable (no misclassification, and a good compromise between the two other criteria).

We noticed however, that without further constraints, as we are fitting averaged data, some of the resulting solutions could induce great variability of behaviour within a group, leading to misclassification. For example, a simulated rat classified as a GT by its parameters could have behaved as a ST and went undetected as its behaviour would have been diluted in the averaged behaviours of the simulated GT group.

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

The fitness function was extended with a new criterion based on the Index Score (see Table 6.2), to favour sets of parameter values that lead to groups of rats that did not introduce such errors, hence without strong inter variability. This is consistent with experimental data [Mey+12]. The resulting new sets of parameter values (see Table 6.3) did not affect the explanatory power of the model.

Objective	Formula
Best fit magazine engagement	$\min(\sum_{s_i \in \text{sessions}} (p^{s_i}_{Sim}(engM\|Group) - p^{s_i}_{Obs}(engM\|Group))^2)$
Best fit lever engagement	$\min(\sum_{s_i \in \text{sessions}} (p^{s_i}_{Sim}(engL\|Group) - p^{s_i}_{Obs}(engL\|Group))^2)$
Penalize parameters that lead to misclassification	$\min(\sum_{a_i \in \text{animals}} \|refPCA(Group) - IndexScore^{a_i}(Group)\|)$

Table 6.2: Revised fitness function. *Lists of the multiple objective/criterion of the fitness function applied to each simulated group.* refPCA *is 1, 0 and -1 for STs, IGs, GTs respectively. This function is combined with NSGA-II to retrieve parameter values that best reproduce the experimental results. It results in a Pareto front of parameters from which we select by hand the solution that is consistent (no agent being misclassified) and that best visually fits the observed data (between engaging with the lever versus engaging with the magazine).*

Type	ω	β	α	γ	u_{ITI}	$\mathcal{Q}_i(s_1, L)$	$\mathcal{Q}_i(s_1, \emptyset)$	$\mathcal{Q}_i(s_1, M)$
ST	0.501	0.243	0.027	0.946	0.845	0.263	0.272	0.344
IG	0.095	0.241	0.885	0.989	0.840	0.059	0.142	0.732
GT	0.081	0.063	0.033	0.483	0.893	0.936	0.022	0.099

Table 6.3: Parameters used to produce the presented results. *All results were generated based on the same parameters. Some parameters might not be used or erased depending on the specific experimental protocol simulated.*

This metric ensures, for example, that using a set of parameter values for sign-tracking will produce a sign-tracker when applying the model in a simulation reproducing the original experiment. Interestingly, it allows us to predict qualitatively what the behaviour of such a rat (ST in normal conditions) would be in new experimental conditions: for example, whether the acquisition or the expression of the behaviour would be blocked or shifted to intermediate or even a goal-tracking behaviour, according to the Index Score defined above.

Note that initial $\mathcal{Q} - values$ have no impact on behaviours on the long run as they are revised by incremental learning during the simulation. Estimated β parameters are sufficient to generate exploration and avoid being permanently biased by such initial values. They mainly help in reproducing the initial tendencies of rats to interact with the experimental environment. They can reflect difference in traits (e.g. novelty-seeking traits) that seem to differ between STs and GTs.

Results

The model has already been validated on a set of behavioural, physiological and pharmacological data [Les+14b]. Interestingly, while the model was only tuned to fit the

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

behavioural data for each group, simulations of additional experiments without changing the parameters were consistent with the remaining experimental data.

The model accounts for the respective engagements of STs and GTs towards distinct specific features [Fla+07; Fla+09; Fla+11b]. It reproduces the difference in patterns of dopaminergic activity for GTs and STs [Fla+11b]. It also reproduces behaviours indicative of incentive salience attribution, including the conditioned reinforcement effect of the lever shown to a greater extent in STs than GTs [RF09], and the consumption-like engagement of the lever or magazine [MB09; DB12b]. Finally, it also reproduces the impact of flupentixol injected either systemically prior to training [Fla+11b], i.e. during acquisition, or locally after the rats have acquired their respective conditioned responses [SR12], i.e. expression.

In the following sections, taking inspiration from the set of studies used to validate the model, we generate predictions that new experiments or extended analyses of the data could confirm.

Dopaminergic patterns of activity

The model parallels the dopaminergic activity recorded in the core of the nucleus accumbens by Fast Scan Cyclic Voltammetry with the RPE signal used in the FMF system. At US time, the RPE signal within the FMF system comes from the difference between the value of the previously engaged cue and the value of the delivered food. At CS time, it mainly reflects the value of the most rewarding cue between the lever and the magazine.

STs and GTs dopaminergic patterns at CS and US time are very distinct [Fla+11b]. While we observe a clear propagation of the signal from US to CS in STs (as expected from the classical RPE theory [Sch98]), this is not the case for GTs for which the CS and US signals are similar to one another and remain relatively constant across sessions (hence, in discrepancy with the classical theory).

In the model, the RPE signal is dependent of the feature previously focused on by the simulated rat. Thus, RPE patterns, averaged over sessions, strongly depend on the dominant path taken by the simulated rats before food delivery. Simulated STs, that mainly engage with the lever before food delivery, have an averaged signal that propagates from US to CS. This reflects that any rat that engages with the lever, eventually learns that it is a full predictor of food delivery. Simulated GTs, that mainly engage with the magazine before food delivery, have an averaged signal that do not show such a propagation. Indeed, the magazine is not fully informative of food delivery for any rat, hence a persistent reward prediction error remains at food delivery when engaging with the magazine during CS.

In Flagel et al. [Fla+11b], recordings of dopaminergic activity in outbred rats were made to parallel those of the selectively bred STs and GTs but no recordings were made in outbred IGs. We would expect that IGs, whose behaviour fluctuate between sign-tracking and goal-tracking, would have a kind of mixed signal, averaging between those following from sign-tracking and goal-tracking. The current parameters values used in the model suggest that we would expect a high signal at CS time that would converge to a certain point, while at the meantime, the signal at US time would keep fluctuating

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

without fully disappearing (see Figure 6.2).

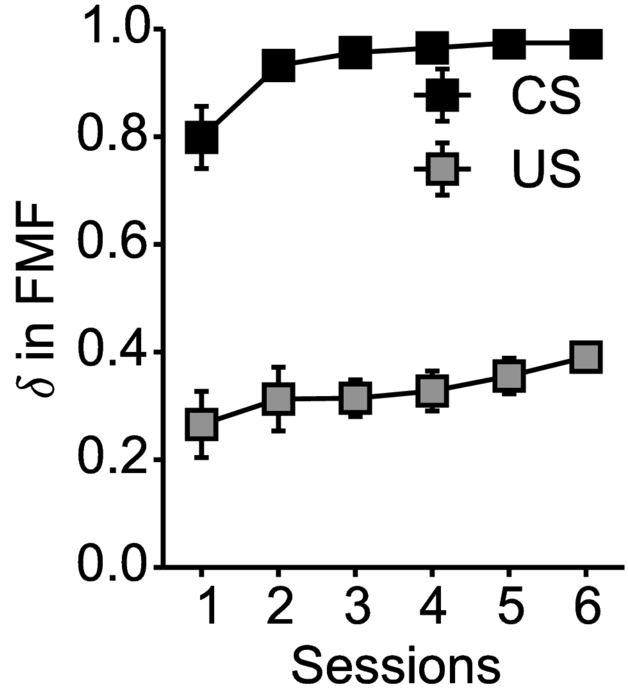


Figure 6.2: Prediction of the model about expected patterns of dopaminergic activity in intermediate rats. *Data are expressed as mean ± SEM. Average RPE computed by the FMF system in response to CS and US presentation for each session of conditioning in the intermediate group.*

Note that the visual results of this prediction are not identical with those in Lesaint et al. [Les+14b]. Contrary to ST and GT behaviours that deeply rely on the mechanisms, IG results strongly depend on the parameter values, which are significantly different with the introduction of the new score. Experimental recordings could help us refine the appropriate set of values for further predictions.

The initial analysis [Fla+11b] and its reproduction [Les+14b] was done without taking into account the features engaged by animals prior to food delivery, possibly averaging very distinct patterns. The model predicts that if we were to organize the data per groups and actions rather than only per groups, we would observe patterns as shown in Figure 6.3. At the time the CS is presented, there should be no differences as all rats are exploring the world and not expecting the lever appearance, hence the positive RPE common to all rats. The difference would be at US time.

STs previously engaged with the lever (Figure 6.3 A) would show a classical propagation pattern, similar to the one of the initial analysis, as this condition dominates in the data. It reflects the fully predictive value of the lever. STs previously engaged with the magazine (Figure 6.3 C) would show a significant peak of DA activity, as they almost never engage with the magazine and hence attribute a low value to it, leading to an expected significant RPE.

GTs previously engaged with the magazine (Figure 6.3 D) would show an absence of propagation and patterns of DA activity that follow those at CS time, similar to the one of the initial analysis, as this condition dominates in the data. It reflects the difference between the value of the food delivered and the lower motivational value of the magazine. GTs previously engaged with the lever (Figure 6.3 B) would show a noisy dopaminergic activity that would decrease with time as the predictive value of the lever is learn.

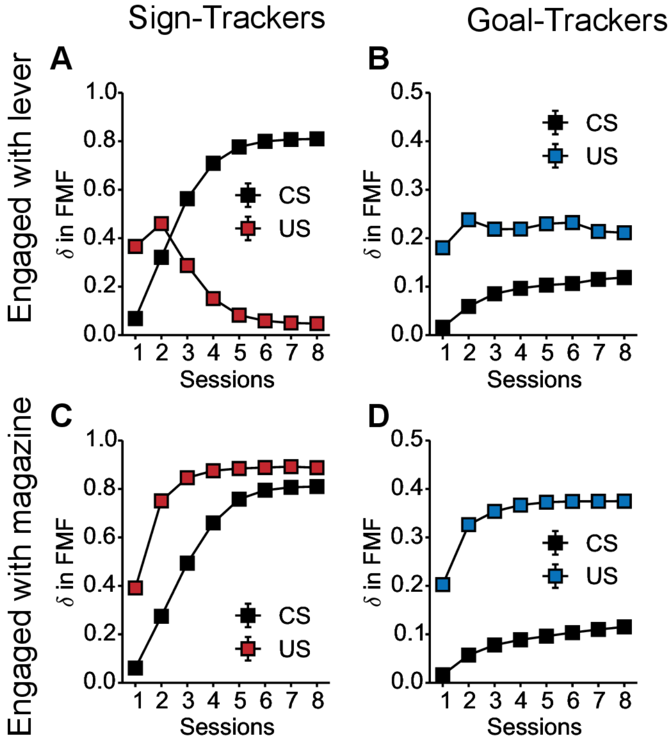


Figure 6.3: Predictions about patterns of dopaminergic activity per groups and per actions. *Average RPE computed by the FMF system in response to CS and US presentation for each session of conditioning for STs* **(A,C)** *and GTs* **(D,B)** *when engaged with the lever* **(A,B)** *or with the magazine* **(C,D)***. The model predicts very distinct patterns of activity depending on the feature engaged with prior to food delivery.*

Removal of magazine during the ITI

In the present model, the simulation of the ITI has a significant impact on the data. We hypothesize that the permanent presence of the magazine during the whole experiment lead animals to revise its associated motivational value, upward at lever retraction (i.e. food delivery) and downward during the ITI as there is no reward to be found then. Hence, on average, its presence does not guarantee access to food. In contrast, the time-locked presence of the lever before food delivery would lead to learn and maintain the motivational value of the lever to a certain level, as its presence guarantees food to be delivered.

First, by keeping the motivational value of the lever higher than that of the magazine in the FMF system, it makes simulated rats favouring this system (STs) to follow a sign-tracking policy. The small contribution of the MB system, which would attract rats towards the magazine does not compensate. Thus, the presence of the magazine in ITI is central for the emergence of STs in the model.

Second, by revising downward the magazine value between episodes, it maintains a discrepancy between the expectation (value) and the observation (reward) at food delivery in simulated rats being engaged with the magazine. This leads to the persistent positive RPE at US time and prevents a full propagation of the signal to CS time. Thus, the presence of the magazine in ITI is also central for the model, to explain the distinct dopaminergic patterns of activity in STs versus GTs that have been observed in Flagel et al. [Fla+11b].

Third, we also hypothesize that values of the FMF system account for the motivational engagement, i.e. incentive salience, observed in rats towards either the lever or the magazine. The higher motivational value of the lever relative to that of the magazine implies that simulated rats chew/bite more the lever than the magazine. While not central to the model, it is consistent with experimental observations [MB09; DB12b].

If no magazine were available during the ITI then, according to the model, the magazine would not loose its motivational value, as it would become a full predictor of food delivery and be highly valued. Hence, we would expect (1) an increased motivational engagement (chew/bite) towards the magazine, (2) a decreased tendency in sign-tracking within the population and (3) a different pattern of dopamine activity when goal-tracking for all rats.

As the motivational value of a feature accounts for the level of motivational engagement towards it, a higher motivational value of the magazine, relative to a control group, would necessarily lead to a relatively stronger motivational engagement towards it.

As the motivational value of the magazine would be as high as that of the lever, there should be no reason for rats relying mainly on the FMF system (STs) to favour one over the other, hence shifting to behaviours similar to those of IGs and GTs (see Figure 6.4). GTs, relying mainly on the MB system would not be deeply affected (see Figure 6.4 B).

Finally, as the presence of the magazine would be time-locked to the moments before the delivery of food, we would expect a propagation of the dopamine signal from US time to CS time (see Figure 6.5). At some point (after the value of the food has been fully learned) the signal at US time should start decreasing. Note that if we would have used the same parameters (except for the weighting parameter) to simulate STs and GTs, we would have expected an identical RPE signal for STs and GTs, and we know this is not the case based on existing data [Fla+11b].

The expected decreased tendency in sign-tracking within the simulated population does not mean that simulated rats would not be attracted any more by the lever. Simulated rats would indeed be attracted by both the lever and magazine because their FMF system attributes a high motivational value to all signs preceding reward delivery. Combined with the contribution of the MB system which attracts rats towards to magazine, it could make

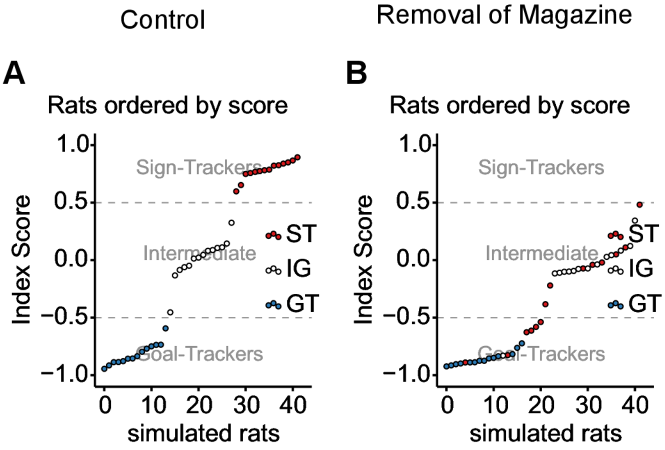


Figure 6.4: Distribution of rats given the removal of the magazine during the ITI. *Simulated rats ordered by their Index Score. In blue simulated rats using parameter values tuned for GTs, in red for STs and in white for IGs in the classical condition (control). Rats with a score* < -0.5 *are GTs, with a score* > 0.5 *are STs and remaining rats are IGs.* **(A)** *As expected, rats using parameters' value for GTs are classified as GTs. Same for STs and IGs.* **(B)** *Without magazine during the ITI, simulated rats that would have been classified as GTs in normal conditions are still classified GTs. However, rats that would have been classified as STs (red) have a score that classify them as GTs or IGs. One IG (white) is now classified as GT.*

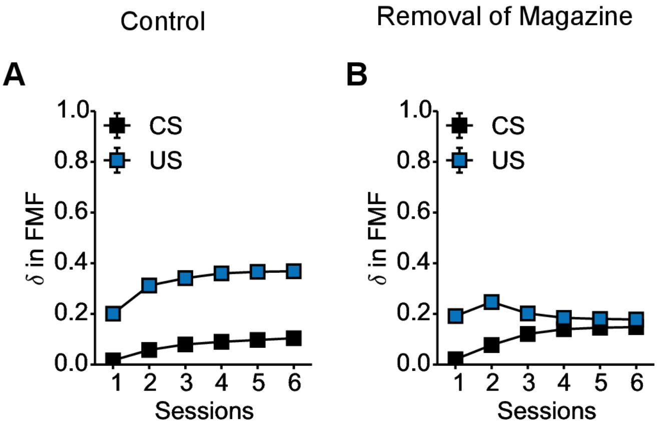


Figure 6.5: Patterns of dopaminergic activity for GTs given the removal of the magazine during ITI. *Average RPE computed by the FMF system in response to CS (black) and US (blue) presentation for GTs for each session of conditioning. It is hypothesized to parallel the patterns of dopaminergic activity observed by FSCV in the core of the nucleus accumbens.* **(A)** *With the classical protocol (control), signal at CS and US seems to follow similar trends and there is no propagation of signal from US to CS.* **(B)** *When magazine is time-locked to CS presentation, the value of US is propagated to the CS. Thus, the signal at US time, after sufficient learning (2 first sessions) start decreasing in favour of the CS time.*

Reprinted from Journal of Physiology - Paris, Florian Lesaint, Olivier Sigaud, Jeremy J. Clark, Shelly B. Flagel and Mehdi Khamassi, Experimental predictions drawn from a computational model of sign-trackers and goal-trackers, in press (doi:10.1016/j.jphysparis.2014.06.001), Copyright (2014), with permission from Elsevier.

the simulated animal engage more with the magazine than with the lever. Thus if the computational model is valid, this would mean that the tendency to sign-track in real animals can be gradually changed by affecting some of the signs or features present in the context of the task (here the magazine during the ITI).

Injections of flupentixol in the core of the nucleus accumbens

In the model, flupentixol, an antagonist of dopamine, is hypothesized to impact the RPE (hypothesized to parallel phasic dopamine) used in the FMF system, putatively based within the core of the nucleus accumbens. Flupentixol is also assumed to affect any action selection process, relying on tonic dopamine [Hum+12]. Hence, under systemic injections of flupentixol, the learning process of the FMF system is disrupted and actions are almost randomly picked barely using learned values.

With systemic injections of flupentixol [Fla+11b], no goal-tracking nor sign-tracking is expressed in the population. However, when afterwards released from flupentixol, GTs fully express goal-tracking, whereas STs behave as untrained rats.

The model accounts for the absence of behaviours under flupentixol by the hypothesized impact of flupentixol on the action selection process, blocking the expression of any acquired behaviour Lesaint et al. [Les+14b]. The subsequent absence of sign-tracking on a last session free of flupentixol is explained by the disruption of the FMF system during the 7 first sessions, blocking behaviour acquisition. The full expression of goal-tracking as soon as flupentixol is removed, relies on the unaffected learning process in the MB system, assumed to be dopamine-independent and hence keeps learning under flupentixol, but which values are simply not used by the softmax function.

The model predicts that if flupentixol were injected locally in the core of the nucleus accumbens rather than systemically prior to acquisition, GTs would normally express their behaviour, as the action selection mechanism would not be disrupted and make use of the values learned in the MB system; whereas STs' behaviour would remain blocked because of the disruption of the FMF system (see Figure 6.6), and this is indeed what happend when Saunders and Robinson [SR12] locally injected flupetixol after the behaviours were already acquired.

Lesions of core of the nucleus accumbens

While we did not try to find all anatomical counter parts of the mechanisms involved in the model, the hypothesis that the FMF system relies mainly on the core of the nucleus accumbens is important for the model. Indeed, RPEs used in the FMF system are compared with the dopaminergic recordings (using FSCV) in the core of the nucleus accumbens. As already stated, the values learned by the FMF system are a key component in the emergence of sign-tracking behaviours within a population and assumed to reflect the motivational engagement observed towards the magazine and the lever.

As stated in the previous section, Flagel et al. [Fla+11b] studied the impact of systemic injection of flupentixol on the acquisition of sign-tracking and goal-tracking. They

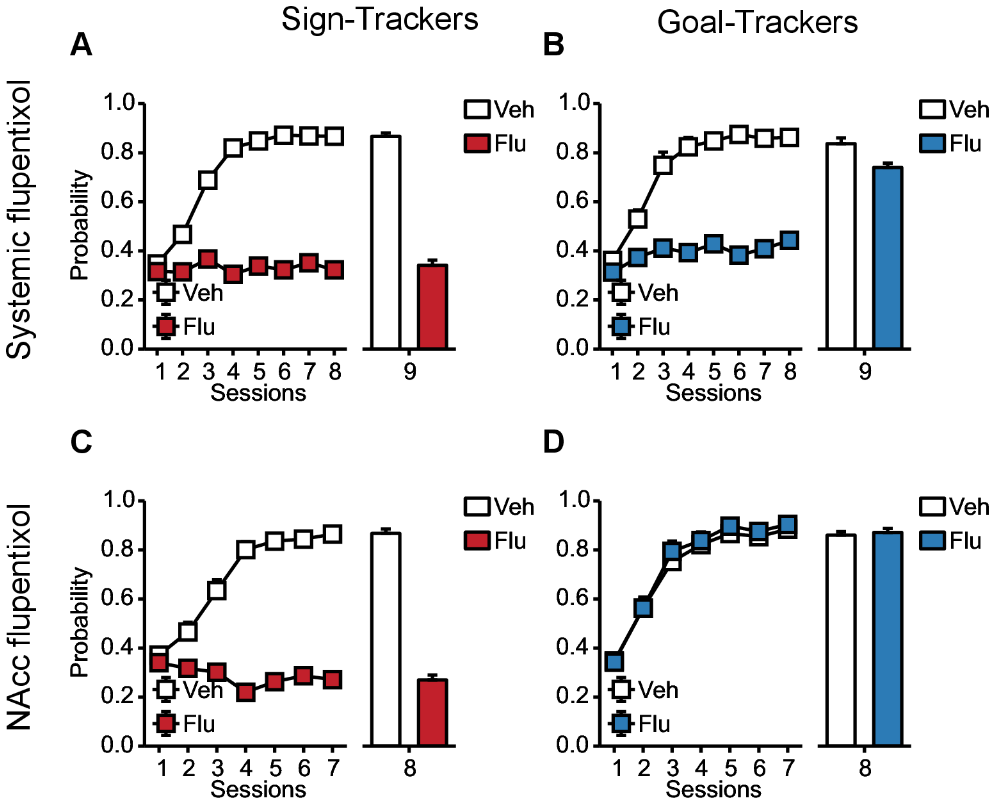


Figure 6.6: Simulated impact of injections of flupentixol. **(A,B)** *Simulation of the impact of systemic injections of flupentixol [Fla+11b] on the probability to approach the lever for STs* **(A)** *and the magazine for GTs* **(B)** *during lever presentation. Last session is without flupentixol. Under flupentixol (first 7 sessions), both sign-tracking and goal-tracking are blocked. On the last flupentixol-free session (8 session), STs are unable to express sign-tracking, its learning having been blocked, whereas GTs fully express goal-tracking, which learning was only covert.* **(C,D)** *Simulation of the impact of local injections of flupentixol in the core of the nucleus accumbens, hypothesised to impact only the FMF-system. Contrary to the initial experiment, the injections being localized to the FMF-system, the action selection mechanism is not impacted. Hence, GTs fully express goal-tracking during the first 7 sessions* **(C)**. *STs are still unable to express sign-tracking* **(D)**.

observed that the acquisition of a goal-tracking behaviour did not require a fully functional dopaminergic system contrary to sign-tracking. Another study [SR12] focused on the impact of local injections of flupentixol in the core of the nucleus accumbens on the expression of sign-tracking and goal-tracking, after 8 days of conditioning. On the last day, with a sufficient dose of flupentixol, they observed a decrease in the general tendency to sign-track in the overall population while leaving the level of goal-tracking unaffected.

Simulating injections of flupentixol in the core of the nucleus accumbens, by disrupting RPEs in the FMF system and hence its contribution in the decision, the model accounts for these last observations. The action selection mechanism remains functional and makes use of the MB system values, such that the behaviour of GTs is preserved while the one of STs is disturbed and leads to a decrease in sign-tracking in the overall population.

We expect that lesions of the core of the nucleus accumbens would lead to similar effects as the above experiments.

Lesions of the core of the nucleus accumbens *prior* to the experiment would (1) block the expression of sign-tracking responses and (2) stop the motivational engagement towards the magazine or the lever during approaches.

By disabling the FMF system (setting and keeping all values to 0), it cannot favour the lever over the magazine any more. STs would therefore act randomly , approaching lever and magazine indifferently, as observed in IGs. We would expect a shift towards goal-tracking similar to the one expected for removing the magazine during the ITI (as in Figure 6.4).

However, while a magazine removal would lead to an increase in motivational engagement, we expect such a lesion to block any consumption-like behaviour. Especially, we would expect GTs' approach behaviour to remain similar to control group, but without subsequent chewing and biting of the magazine.

We would expect that lesions of the core of the nucleus accumbens *after* the experiment would disrupt the tendency to sign-track in the overall population, while leaving the tendency to goal-track intact (see Figure 6.7). However, contrary to flupentixol injections, that needed 35 min of infusion for a visible effect, we would expect the effect to be immediate with a lesion. Such a lesion would disrupt the FMF system, hence (1) suppressing any consumption-like engagement towards the features (motivational values being kept to 0), and (2) stop favouring engagements towards the lever. The lesion would leave the MB system unaffected and have no impact on the general tendency to goal-track.

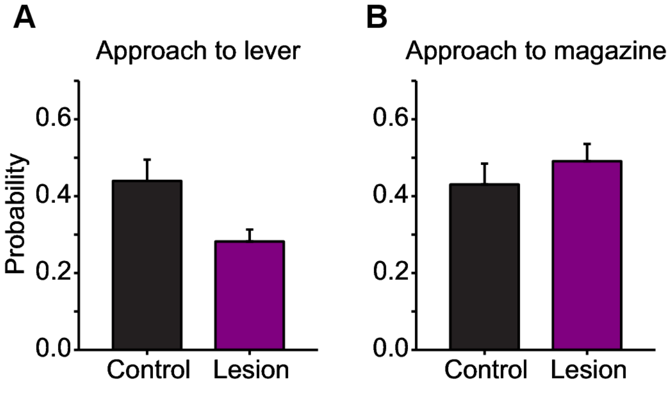


Figure 6.7: Predictions of the impact of lesions of the core of the nucleus accumbens after conditioning. *General tendencies to sign-track* **(A)** *and goal-track* **(B)** *in a population of rats after training. Lesion of the core of the nucleus accumbens is simulated by a blockade of the FMF system in the model. We expect a decrease in the tendency to sign-track* **(A)** *with a lesion (purple) relative to a control group (black). General goal-tracking tendencies should remain unchanged* **(B)**.

Discussion

Relying on a model that was previously validated using experimental data to account for variability in rats undergoing an autoshaping paradigm [Les+14b], we generate an additional set of behavioural, physiological and pharmacological predictions.

We predict that dopaminergic patterns for IGs should be a mixed signal between those observed for STs and GTs. We predict that looking separately at the DA patterns given the prior engagement towards either the lever or the magazine should lead to clearly distinct patterns. We predict that the removal of the magazine during the ITI should lead to an increased motivational engagement towards the magazine, a decreased tendency in sign-tracking within the population and a different pattern of dopaminergic activity when goal-tracking. Finally, we predict that local injections of flupentixol to the core of the nucleus accumbens would preserve goal-tracking and prevent the learning of a sign-tracking response, a result that should also be observed following lesions of the core of the nucleus accumbens prior to conditioning. Lesions after conditioning, would only block the expression of the learned sign-tracking behaviour.

An important limitation of the present predictions is that most of them are based on the behaviour that is expected to emerge from naive rats trained in a revised protocol, assuming that they would have behaved in a specific manner in the standard protocol (e.g. expecting a supposed ST to goal-track). To overcome this difficulty, one must look at the population level rather than the individual level [SR12], which might be problematic as the proportion of GTs, STs and IGs is highly variable in a population [Mey+12; Fit+13]. An alternative would be to use selectively bred rats that can more or less be ensured to behave as STs or GTs in experimental conditions [Fla+11b].

Another limit of the present predictions are the hypotheses on which they are based. It cannot be excluded that the core of the nucleus accumbens also contributes to the MB system, but not by its dopaminergic activity [KH12; MR11; McD+11] (but see Meer et al. [Mee+10]; Bornstein and Daw [BD12]; Penner and Mizumori [PM12]). Hence, completely disrupting it might unexpectedly affect goal-tracking. Validating these predictions would help to confirm this hypothesis. In the initial model [Les+14b] we interpreted the parameter which simulates the ITI as accounting for the engagement of the rats towards the magazine during the ITI. Preliminary analyses of experimental data (not shown), while still inconclusive, tend to mitigate such a strong hypothesis. Hence, in the current article, we only assume that the presence of the magazine during the ITI impacts its general motivational value within the experiment. Validating such predictions would definitely help to clarify the impact of the ITI context on the expressed behaviours.

One could argue that, to some extent, describing STs with a MF system and GTs with a MB system could be sufficient to explain dominant behaviours [Cla+12]. However, it would fail to explain the full and continuous spectrum of observed behaviours [Mey+12]. If the predictions that we make about IGs (which have an intermediate behaviour between STs and GTs) are correct, this would argue in favour of a continuum in the weighting between MB and MF systems rather than a pure dichotomy.

An alternative to the collaboration of both systems (through a weighted sum) would be a reciprocal inhibition, such that only one system would be working at a time. This would be sufficient to account for the previous point and may even be able to account for the absence of RPE pattern in the dopamine signal measures in GTs [Fla+11b] without requiring a revision of the magazine value during ITI. The inhibition of the MF system in GTs would indeed prevent any RPE signal from being observed. However, it would be unable to properly account for the consumption-like engagement observed in both STs and GTs without some kind of extension (see Zhang et al. [Zha+09] for a computational model of incentive salience). It would also fail to explain why the pharmacological disruptions of one system does not seem to let the other take control [Fla+11b].

Another possibility would be that the two systems run in parallel but that only one is used to make the decision during a trial. Assuming that one system leads to the lever and the other to the magazine, we would expect IGs to behave as STs when engaging with the lever and GTs when engaging with the magazine. Experimental data goes against such interpretation. Meyer et al. [Mey+12] observed that contrary to STs or GTs, IGs tend to approach both the magazine and the lever during single trials. Some rats even hold on to the lever while putting their head into the magazine (which no model that selects a single action at a time can reproduce). While the task representation does not allow multiple engagements in a trial, this suggests that both systems are active and contribute actively to their behaviour at all time. We would also expect rats to behave differently when using one system over the other, such that, for example, rats would actively engage with the lever but quietly wait in front of the magazine, which is not the case. Finally, the recent literature seems consistent with multiple systems working in parallel and partially contributing to a global decision (e.g. Daw et al. [Daw+11]). Hence, this does not suggest take-over competition between the systems. Trial-by-trial analyses [Daw11] would allow us to definitely rule out such alternatives. Finally, if only the output of the MF system was inhibited, given that the lever appearance is fully predictive of food delivery, no classical MF system (relying on classical state representation) would reproduce the differences observed in phasic dopaminergic patterns between STs and GTs nor explain the differences of focused features. Hence, the model suggests to take features into consideration.

The interest of the current computational model lies in its combination of simple concepts actively used and accepted in the current field (Dual reinforcement learning and factored representations) but rarely used together, to account for a variability of experimental data, without resorting to arbitrary additions. As a result the current model does not behave as state of the art algorithms would on the same task and produces a suboptimal behaviour. This suboptimal behaviour is, however, in accordance with behavioural observations in rats.

Subsequent studies could benefit from a different approach to estimate parameters. We are currently fitting the model on the behavioural data per sessions and groups, using trial-by-trial analyses could prove a better tool to fit the parameters at the individual level [Daw+11] and comfort some choices in the architecture of the model.

It has been suggested that individuals for whom cues become powerful incentives (i.e. STs) are more prone to develop addiction [SR12]. Thus, the current model and its predictions will allow us to further investigate and possibly identify the neural mechanisms that underlie addiction and related disorders. For example, the current model predicts that some manipulations could alter the behaviour of STs towards that of GTs, and the neurobiological targets of these manipulations may be used to alter drug-cue dependency and prevent relapse (For further discussion regarding the role of learning-related dopamine signals in addiction vulnerability, see Huys et al. [Huy+14]).

To conclude, the current article refines the model previously described by Lesaint et al. [Les+14b] with an additional metric that strengthens its explanatory power. It mainly suggests a set of predictions with which to further confront the model. The new proposed experiments would help to better localize the anatomical counterparts of the mechanisms involved and disentangle their contributions to the observed behaviours. It would also help in refining the hypotheses and simplifications of the model and we hope would confirm the interest and necessity of considering the features rather than the general situations encountered by rats when modelling this kind of phenomena.

Chapter 7

This chapter presents an application of the computational model developed in Chapter 5 to a new set of experimental data. This work is presented under the form of a journal paper currently under review:

> *Florian Lesaint, Olivier Sigaud, and Mehdi Khamassi. "Accounting for negative automaintenance in pigeons : A dual learning systems approach and factored representations". In:* PLoS One *(under review)*

It aims at investing the opportunity to use dual-learning systems combined with factored representations [Les+14b] to study not only Pavlovian conditioning but also Pavlovian and instrumental interactions. It once again proves the interest of such a method by confronting the model to experimental data that have not been much accounted for by computational neuroscientists, especially as it provides an explanation for conflicting results in the current literature.

It shows that the computational model can reproduce inter-individual differences in pigeons undergoing a negative automaintenance task [WW69; San+06]. It also suggests some predictions that could confirm or revoke the model if tested in subsequent studies. Simulations suggest that the variability of behaviour in pigeons can be interpreted in a way similar to that of rats, that is by the presence of sign-trackers and goal-trackers, which differently rely more on one system than on the other. The model also explains some additional properties of the behaviours investigated in [WW69].

Accounting for Negative Automaintenance in Pigeons: A Dual Learning Systems Approach and Factored Representations

Florian Lesaint^{1,2,*}, Olivier Sigaud^{1,2}, Mehdi Khamassi^{1,2}

1 Institut des Systèmes Intelligents et de Robotique, UMR 7222, UPMC Univ Paris 06, Paris, France

2 Institut des Systèmes Intelligents et de Robotique, UMR 7222, CNRS, Paris, France

$*$ E-mail: Corresponding lesaint@isir.upmc.fr

Abstract

Animals, including Humans, are prone to develop persistent maladaptive and suboptimal behaviours. Some of these behaviours have been suggested to arise from interactions between brain systems of Pavlovian conditioning, the acquisition of responses to initially neutral stimuli previously paired with rewards, and instrumental conditioning, the acquisition of active behaviours leading to rewards. However the mechanics of these systems and their interactions are still unclear. While extensively studied independently, few models have been developed to account for these interactions. On some experiment, pigeons have been observed to display a maladaptive behaviour that some suggest to involve conflicts between Pavlovian and instrumental conditioning. In a procedure referred as negative automaintenance, a key light is paired with the subsequent delivery of food, however any peck towards the key light results in the omission of the reward. Studies showed that in such procedure some pigeons persisted in pecking to a substantial level despite its negative consequence, while others learned to refrain from pecking and maximized their cumulative rewards. Furthermore, the pigeons that were unable to refrain from pecking could nevertheless shift their pecks towards a harmless alternative key light. We confronted a computational model that combines dual-learning systems and factored representations, recently developed to account for sign-tracking and goal-tracking behaviours in rats, to these negative automaintenance experimental data. We show that it can explain the variability of the observed behaviours and the capacity of alternative key lights to distract pigeons from their detrimental behaviours. These results confirm the proposed model as an interesting tool to reproduce experiments that could involve interactions between Pavlovian and instrumental conditioning. The model allows us to draw predictions that may be experimentally verified, which could help further investigate the neural mechanisms underlying theses interactions.

Introduction

Persistent maladaptive and suboptimal behaviours are commonly observed in animals, including Humans, and supposed to results from possible constraints (e.g. energy versus efficiency trade-off) solved by the interaction of neural mechanisms not clearly identified yet. Breland and Breland [BB61] studied animals that learned to retrieve rewards given some action (e.g. drop an object to get food). They observed that, while successful at first, these animals developed strange behaviours which blocked them in achieving the rewarding action (e.g. paws kept clenched on the food-predicting object). Hershberger [Her86] studied how chicks failed to learn to run away from visible food to eventually get access to it. Guitart-Masip et al. [GM+12] showed that many humans have difficulties to learn to withhold from acting to get rewarded in a go/no-go task. These maladaptive behaviours have been suggested to arise from the interactions between multiple decision systems in the brain [Day+06; Red+08; BD10; Cla+12], namely Pavlovian and instrumental systems. Pavlovian conditioning is the acquisition of responses associated to initially neutral stimuli that have been paired with rewards while instrumental conditioning is the acquisition of an active behaviour in order to retrieve rewards or avoid punishments. However, the respective mechanisms of these two types of conditioning and how they interact are still unclear.

An example of such maladaptive behaviour was experimentally investigated by Williams and Williams [WW69], whose initial goal was to explore the properties of the pecks developed by pigeons in procedures subsequently referred as *autoshaping* [Ski38]. A classical autoshaping procedure elicits a standard Pavlovian phenomenon. It consists in pairing a conditioned cue (e.g. a light) with the subsequent delivery of food and results in animals developing robust conditioned responses (e.g. pecks) towards the conditioned cue, even if these responses were unnecessary to be rewarded. Actually, Brown and Jenkins [BJ68] found autoshaping to be a more effective way of getting animals to engage with objects for subsequent instrumental experiments, such as pulling a chain or pressing a lever, than other training protocols. Williams and Williams [WW69] developed another protocol, that was afterwards referred as a *negative automaintenance procedure*, which consisted in a setup identical to an *autoshaping* procedure, with the exception that pecking the light turned it off and reward was subsequently omitted. Unexpectedly, they observed that most of their pigeons persisted, although to a lower extent, to peck the light despite its negative consequence, losing during the process a significant amount of reward. The phenomenon was further investigated in both pigeons [DW77; GR73; Kil03; Woo+74], and other species such as rats [Loc+76; Loc+78; O'C79] and rabbits [GH72] with similar results. However, in a more recent study on pigeons with a slightly different negative automaintenance procedure, Sanabria et al. [San+06] did not observe as much sustained detrimental pecks as observed by Williams and Williams [WW69], casting a shadow over the original results. While the differences in the procedures might be one reason of such conflicting results, the present paper develops an additional possible reason.

According to multiple studies [Day+06; Loc+78; San+06], negative automaintenance investigates the confrontation between Pavlovian processes and instrumental ones. It is our interpretation that conditioned responses develop because of the contingency between the conditioned stimulus and the reward (Pavlovian conditioning) and one would expect pigeons not to peck as it prevents them from being rewarded (instrumental conditioning). Understanding the underlying neural mechanisms that result in such behaviours is also important to clarify the constraints and strategies developed by years of evolutions for animals to survive in nature.

Killeen [Kil03] and Sanabria et al. [San+06] have proposed computational models to account for the pecking behaviour described above. However their models are very specific to the task and not easily generalizable to the study of other phenomena. Dayan et al. [Day+06] proposed a more general computational model of interactions between Pavlovian and instrumental conditioning and took negative automaintenance as an illustration, focusing on the first experiment of Williams and Williams [WW69], that introduces the general phenomenon, but without investigating its subtleties resulting from more specific subsequent experiments.

Initially inspired by this latter model, Lesaint et al. [Les+14b] developed a computational model that accounts for a variety of experimental results in rats undergoing an autoshaping procedure [Fla+11b], especially observed inter-individual variabilities of behaviours within the population. In this study, some rats (sign-trackers) came to approach and engage the conditioned stimulus (CS) itself more and more avidly, whereas other rats (goal-trackers) learned to approach and engage the location of food delivery upon CS presentation, a variability also visible at the physiological and pharmacological level.

In the present study, we show that the model of Lesaint et al. [Les+14b], initially developed to account for autoshaping in rats, can reproduce with barely no modifications the experimental data on autoshaping and negative automaintenance in pigeons. Especially, the model suggests as one of the plausible reasons regarding the conflicting data of Williams and Williams [WW69] and Sanabria et al. [San+06], that the variability of observed behaviours partially results from the presence of sign-trackers and goal-trackers within pigeons. It is also able to account for other experimental data about the necessary properties of the cues to express negative automaintenance [WW69]. Moreover, the model generates predictions that may be tested with additional experiments. We further discuss the interest of the combination of concepts on which the model relies for the reproduction of experimental data on Palovian and instrumental conditioning.

Methods

Model

The model from which the present results are generated is described in depth in [Les+14b]. Here we describe the computational mechanisms of the model that capture the experimental data in pigeons. The model is based on a reinforcement learning (RL) method, which describes how an agent should adapt its behaviour to rewarding events. Reinforcement learning relies on Markov Decision Processes (MDP) where the environment is described as a set of states between which the agent can move by acting (see next section). The model is composed of two distinct reinforcement learning systems that collaborate, through a weighted sum integration of values respectively computed by each system, to select an action at each step of the experiment (Figure 7.1) [Cla+12]. One system favours rational and optimal plans of actions while the other leads to more impulsive choices.

The first system is a model-based (MB) system that learns the long term consequences of actions by estimating an approximate model of the world (a transition function $\mathcal{T}$ and a reward function $\mathcal{R}$) on which to build action plans. The model is sufficient to anticipate the delivery of food subsequently to key lights appearance and therefore the interest of being close to the magazine even before its delivery. It is also sufficient to learn that pecking leads to reward omission and should be avoided. This system produces a goal-directed behaviour [Boa77; DB94]. In our implementation of this Model-Based process, the system infers the advantage ($\mathcal{A}$) of taking each action in each situation from its model, given the classical following formulae:

$$\mathcal{Q}(s,a) \leftarrow \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{T}(s'|s,a) \max_{a'} \mathcal{Q}(s',a') \tag{7.1}$$

$$\mathcal{A}(s,a) \leftarrow \mathcal{Q}(s,a) - \max_{a'} \mathcal{Q}(s,a') \tag{7.2}$$

where the discount rate $0 \leq \gamma \leq 1$ classically represents the preference for immediate versus distant rewards and $\mathcal{Q}(s,a)$ is the expected value of doing action a in state s (it corresponds to the discounted accumulation of rewards expected from that moment if subsequently following the assumed best plan of actions). At each step, the most valued action is the most rewarding in the long run (e.g. approaching the magazine to be ready to consume the food as soon as it appears). Equation 7.1 reflects the prospective process by which the simulated agent estimates the future consequences of performing action a in state s. If action a is assumed to lead to a reward $\mathcal{R}(s,a)$ or with a good probability $\mathcal{T}(s'|s,a)$ to another state s' with a high quality action $\mathcal{Q}(s',a')$ then the agent will associate a high $\mathcal{Q}$-value to the state-action pair $\langle s,a \rangle$. Equation 7.2 deduces the advantage of performing action a in state s by comparing its $\mathcal{Q}$-value with the maximal possible $\mathcal{Q}$-value of all available actions in the same state. Note that other implementations could be possible.

The second system is model-free (MF). It does not learn an internal model of the world but incrementally learns to associate values to features of the environment, favouring actions towards valued ones. As a result, this system produces a reactive behaviour in a way similar to habits [Gra08; DD13]. Without an internal model, it cannot consider the consequences of an action and hence solely bases its decision on the a priori expectation values it learns.

In traditional RL (e.g. the MB system), values are learned over abstract states (e.g. arbitrarily defined as s_1, $s_2 \ldots s_x$), such that similarities between situations (e.g. presence of a magazine) are ignored. The present system learns values ($\mathcal{V}$) over features (e.g. food, lever or magazine) and is further defined as the feature model-free system (FMF). Using features reintroduces the capacity to use and benefit from similarities between states. The incremental learning of values relies on a reward prediction error (RPE) signal δ, and works as follows:

$$\begin{aligned} \mathcal{V}(f) &\leftarrow \mathcal{V}(f) + \alpha\delta \\ \delta &\leftarrow r + \gamma \max_{f' \in s'} \mathcal{V}(f') - \mathcal{V}(f) \end{aligned} \tag{7.3}$$

where f is the feature that has been focused on by the action a in state s. The max suggests that all the features f' of the new state s' are considered and the most valued one is used to compute the RPE, even if it might not be the feature focused by the next chosen action. This update rule (Equation 7.3) may be paralleled with the one of the classical Model-Free $\mathcal{Q}$-Learning algorithm [SB98] where $\mathcal{Q}$-values are used in place of $\mathcal{V}$-values. While very similar, such rules can actually produce very different results and patterns depending on the involved situations. The model embeds a feature-function $c : \mathcal{S} \times \mathcal{A} \rightarrow \{keylight(s), magazine, food, \emptyset\}$ that returns the feature the action a was focusing on in state s (e.g. it returns the key light when the action was to engage with the key light). In [Les+14b] we hypothesized that, similarly to classical model-free systems, δ parallels phasic dopaminergic activity [Sch98]. This signal enables to revise and attribute values, seen as motivational or incentive, to features without the need of the internal model of the world used by the MB system. When an event is fully expected, there should be no RPE as its value is fully anticipated; when an event is positively surprising, there should be a positive RPE [Niv09]. The values learned bias the behaviour towards actions that are directed towards the most motivational features (e.g. engaging with the key light would be biased by the general motivational value of the key light). This might lead to favour suboptimal actions with regard to maximizing rewards (e.g. engaging with the negative key light prevents pigeons from being rewarded). The FMF system models the attraction developed by reward-predicting stimuli in such experiments, i.e. incentive salience [MB09; DB12b; Ber07].

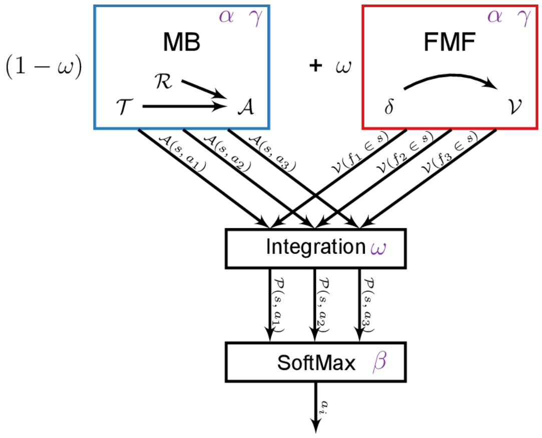


Figure 7.1: Model used for simulations. *The model is composed of a model-based system (MB, in blue) and a Feature-Model-Free system (FMF, in red) which provide respectively an advantage function $\mathcal{A}$ for actions a_i given a state s and a value function $\mathcal{V}$ for each feature f_i that compose the given state. These values are integrated in $\mathcal{P}$, prior to be used into an action selection mechanism. The various elements may rely on parameters (in purple).*

The model does not base its decision on a single system at a time. Rather, the values of the MB system ($\mathcal{A}_{MB}$) and the FMF system ($\mathcal{V}_{FMF}$) are integrated such that a single decision is made at each time step. The values computed by these two systems are combined through a weighted sum and transmitted to a softmax action selection mechanism that converts them into probabilities of selecting actions given a situation (Figure 7.1). The integration is done as follows:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{MB}(s,a) + \omega \begin{cases} 0 & \text{if } a = ngo \\ \mathcal{V}_{FMF}(f) & \text{with } f = c(s,a) \text{ otherwise} \end{cases} \tag{7.4}$$

where $0 \leq \omega \leq 1$ is a combination parameter which defines the importance of each system in the overall model. Pigeons may be modelled with a particular ω value, different ω values producing different characteristics of behaviour. The integration (Equation 7.4) differs from the one suggested by Lesaint et al. [Les+14b] as the tasks presented here introduce the new notion of *refraining from engaging*. We hypothesize that refraining from engaging with a stimulus does not benefit from the FMF bonus associated with such stimulus, hence the $a = ngo$ condition in the second part of the equation. This hypothesis is based on studies of go and no-go learning [GM+12; GM+14] that suggest the presence of a bias for engaging against withholding. Note that this modification could be propagated to the previous studies [Les+14b; Les+14a] without any impact. Indeed, the experiments already accounted for by the model do not require to refrain from acting.

The model incrementally learns from experience at each step. FMF and MB systems are updated according to the action a taken by the full model in state s and the resulting new state s' and retrieved reward r.

Task modelling

Figures 7.2, 7.3 and 7.4 show the MDPs used to simulate the different experiments of Williams and Williams [WW69] and Sanabria et al. [San+06]. We assume that each experimental trial can be simulated with a finite horizon episode, that is by a single run in an MDP with an initial and a terminal state. Furthermore, to comply with the MDP framework, we assume that engagement is necessarily exclusive to one or no stimulus and we do not model time, which is sufficient to replicate the experimental data.

In Experiment 1 (Figure 7.2), the agent starts from an empty state (s_0) where there is nothing to do but explore. At some point the key light is turned on (s_1). The agent can either approach the key light (s_2), approach the magazine (s_4) or keep exploring (s_3,s_6). If close to the key light (s_2), it can either engage with it which ends the trial without reward (s_0), or refrain from engaging until food is eventually delivered (s_5). If close to the magazine (s_4), engaging or not has no impact and leads to food delivery (s_7). Finally, if the agent is far from the magazine (s_5,s_6), it first needs to get closer (s_7) before consuming the food, hence retrieving the only available reward in this trial (R). It ends in an empty state (s_0) which symbolizes the start of the inter-trial interval (ITI): no food, no lever and *an empty but still present magazine*. Paths in red are those that should be favoured by the FMF system, leading to the potentially detrimental action of engaging with the Key light. Paths in blue are those that should be favoured by the MB system, successfully leading to reward delivery.

Experiments 3 and 4 use additional key lights (irrelevant and continuous). Each light extends the previous MDP with an additional path as described in Figures 7.3 and

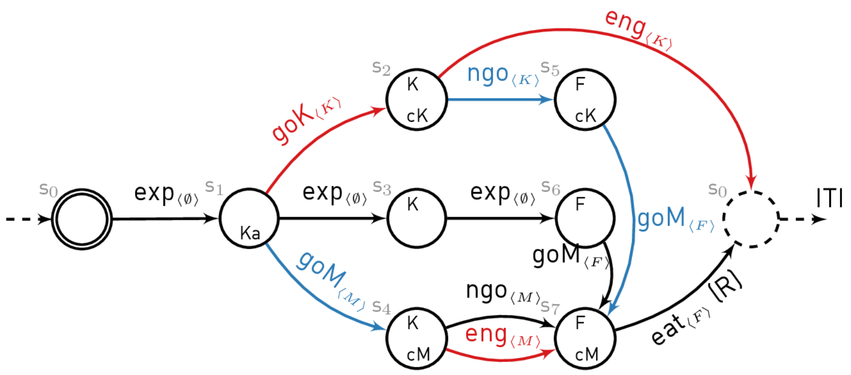

s_0
$\mathrm{exp}_{\langle\emptyset\rangle}$
s_1
Ka
$\mathrm{goK}_{\langle K\rangle}$
s_2
K
cK
$\mathrm{ngo}_{\langle K\rangle}$
s_5
F
cK
$\mathrm{eng}_{\langle K\rangle}$
$\mathrm{exp}_{\langle\emptyset\rangle}$
s_3
K
$\mathrm{exp}_{\langle\emptyset\rangle}$
s_6
F
$\mathrm{goM}_{\langle F\rangle}$
$\mathrm{goM}_{\langle F\rangle}$
s_0
ITI
$\mathrm{goM}_{\langle M\rangle}$
s_4
K
cM
$\mathrm{ngo}_{\langle M\rangle}$
s_7
$\mathrm{eng}_{\langle M\rangle}$
F
cM
$\mathrm{eat}_{\langle F\rangle}$ (R)

Figure 7.2: Computational representation of the negative automaintenance procedure. *MDP accounting for Experiment 1 in Williams and Williams [WW69] and for the Brief PA protocol of Sanabria et al. [San+06]. States are described by a set of variables: K/F - negative Key light/Food is available (Magazine is always available, hence it is not shown), cM/cK - close to the Magazine/negative Key light, Ka - Key light appearance. The initial state is double circled, the dashed state is terminal and terminates the current episode. Actions are engage (eng) or refrain from engaging (*ngo*) with the proximal stimuli, explore (exp), or go to the* Magazine/Key *light and* eat*. Only the* eat *action is rewarded (R), such that in this experiment, pigeons that engage with the Key light receive nothing during the trial. For each action, the feature being focused on is displayed within brackets.*

7.4. The main idea is that animals can orient towards any key light (or magazine) and subsequently engage with it. Based on the simulated protocols, paths can be activated/deactivated during experiments, such that only available actions are considered by the model in its decision. In Experiment 3, the role of the keys (K and I) are reversed multiple times during the experiment (Blocks A and B in Figure 7.3).

In Williams and Williams [WW69], the key light is immediately turned off following a peck. In Sanabria et al. [San+06] protocol, the key light is maintained for a fixed period, whatever the behaviour of the pigeon. Food is then only delivered if no contacts with the key light are made during that period. Pigeons could therefore produce multiple pecks during a trial, hence the difference in scales between both studies that is not replicated in our results. Despite such difference in protocols, the MDP of Figure 7.2 is also used to simulate the results by Sanabria et al. [San+06]. Consequently, we mainly explain the difference of behaviours between the two studies by an inter-individual variability in pigeons, simulated by different parameter values, rather than by the difference in protocols.

Inter-trial interval (ITI)

While the MDP does not model the ITI, we assume that the presence of a stimulus (key light or magazine) during ITI degrades its values in the model. This current hypothesis is simulated by revising the values of the magazine and the continuous key light (if available)

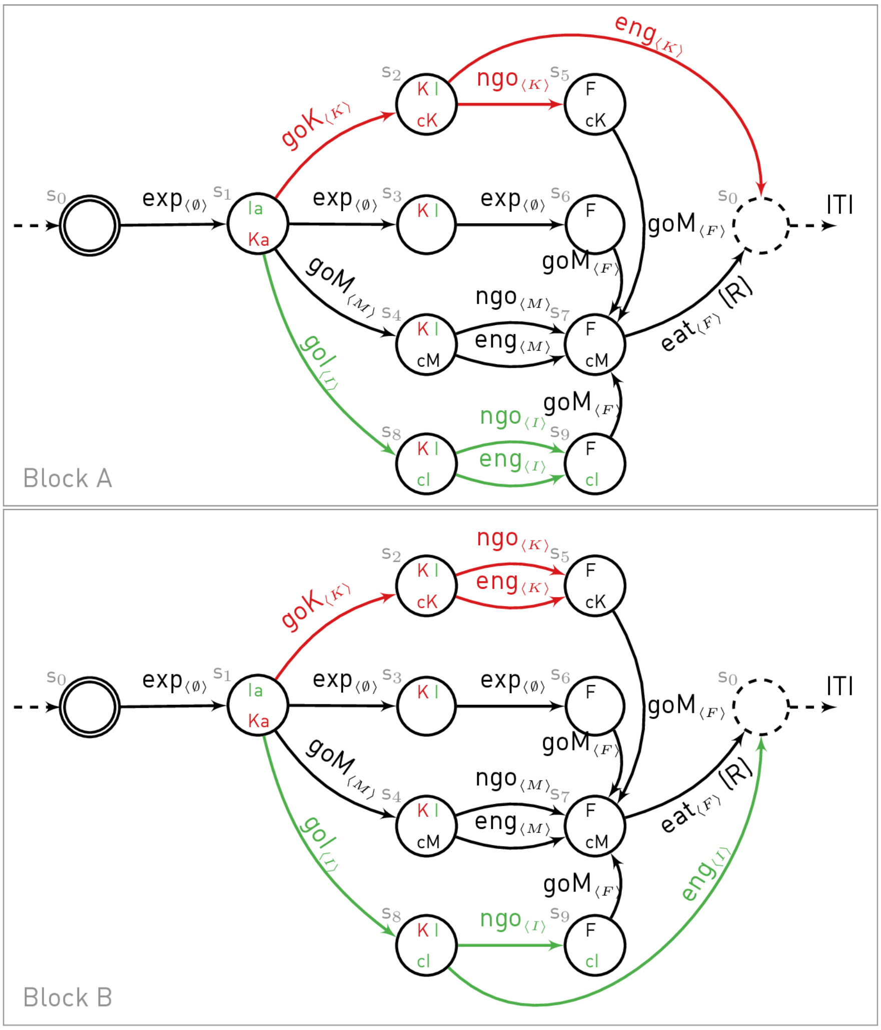


Figure 7.3: *MDP for simulation of Experiment 3 in Williams and Williams.* Legend is as in Figure 7.2. The path involving an engagement with the negative key light is highlighted in red. A new irrelevant key light (green), the associated paths and actions are added to the MDP of Figure 7.3. The animal starts in block A. During the experiment, blocks can be switched without informing the animal, such that the contingencies are reversed between keys.

with the following formulae:

$$\begin{aligned} \mathcal{V}(M) &\leftarrow (1 - u_{ITI}) \times \mathcal{V}(M) \\ \mathcal{V}(C) &\leftarrow (1 - uc_{ITI}) \times \mathcal{V}(C) \end{aligned} \tag{7.5}$$

where the parameters $0 \leq u_{ITI} \leq 1$ and $0 \leq uc_{ITI} \leq 1$ reflect the impact of the presence of the magazine and the continuous key light during ITI on their acquired value in the

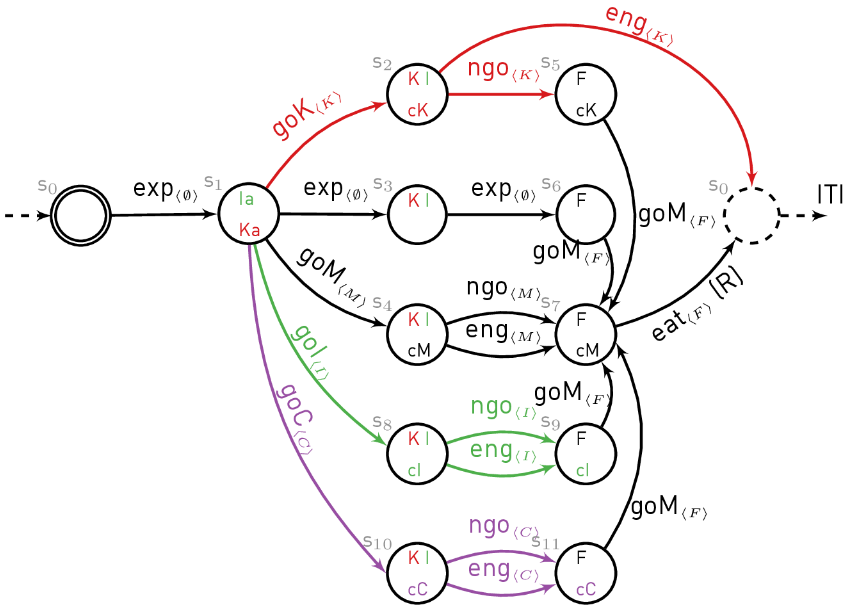


Figure 7.4: MDP for simulation of Experiment 4 of Williams and Williams. *Legend is as in Figure 7.3. A new continuous irrelevant key light (purple), the associated paths and actions are added to MDP of Figure 7.3 (Block A) . Note that while not shown, as for the Magazine, the Continuous key light is present in all states. Paths are activated/deactivated depending on the current phase of the current protocol (Table 7.2).*

FMF system. A low value symbolizes a low impact and therefore a low revision of the value associated to the stimulus.

Note that extending the MDP with a set of states to represent this interval would have increased the complexity of the MDP, introduced non-Markov aspects to the task and increased the time required for simulations. Furthermore, while it might have led to the same results, the interpretation would have been different from our hypothesis, as downgrading the values would have required engagement and not only the presence of stimuli.

Pre-training

No MDP was used to simulate the possible autoshaping pre-training that underwent some of the pigeons in the experiments, nor the necessary familiarization with the Skinner box and the magazine mechanism. Rather, we initialize the model with values ($\mathcal{Q}_i(s_1, goK), \mathcal{Q}_i(s_1, goM), \mathcal{Q}_i(s_1, exp)$) that simulate the action-values acquired during such pre-training phases.

These values have no impact in the long run behaviours as they are revised by incremental learning during the simulation. They mainly help in reproducing the initial tendencies of pigeons to interact with the experimental environment.

Model parameters and simulations

The model relies on a set of 8 parameters (a shared learning rate, a shared discount rate, a selection temperature, an integration parameter and 3 initial conditions) that need to be tuned for simulations to reproduce experimental data. The parameter values used were obtained by hand tuning. More automatic tuning methods (e.g. fitting optimisation algorithms [Les+14b]) were not possible without more precise numerical experimental data. Hence we only tried to qualitatively replicate the experimental results of Williams and Williams [WW69] and Sanabria et al. [San+06].

Nevertheless, simulation results were generated with a single set of parameter values for all experiments of Williams and Williams [WW69] and Sanabria et al. [San+06], with the exception of ω and $\mathcal{Q}_i(s_1, goK)$ (see Table 7.1). Following the terminology used in Lesaint et al. [Les+14b] to categorize rats, we can say that we simulated *sign-trackers* (high ω) and *goal-trackers* (low ω) pigeons.

Varying the ω parameter is sufficient here to reproduce the experimental results. This was done here for parsimony, in order to highlight the key important mechanisms to explain experimental data without giving the model too many degrees of freedom. It is however almost certain that pigeons would not share the exact same parameter values in reality. Especially, breeding procedures, housing procedures and training procedures might have some impact on the averaged neural mechanisms properties modelled with these values.

Sanabria et al. [San+06] pigeons were divided into multiple groups that underwent different protocols, with multiple mixed phases of positive and negative training. Except for 3 pigeons, Williams and Williams [WW69] did not train their pigeons on the key lights before the main experiments. For a better comparison between these studies, we only focus on the pigeons of Sanabria et al. [San+06] that were briefly exposed to autoshaping before being confronted to negative automaintenance (*Brief PA* protocol) and pigeons with no pre-training in Williams and Williams [WW69], hence the difference of value for the $\mathcal{Q}_i(s_1, goK)$ parameter.

Table 7.1: Parameters values used for simulations.

Pigeons	Grp	ω	β	α	γ	u_{ITI}	uc_{ITI}	$\mathcal{Q}_i(s1, goL/goM/exp)$
Williams and Williams *	STs	0.9	0.15	0.2	0.9	0.3	0.2	0.0 / 0.2 / 0.2
Sanabria et al.	GTs	0.2	0.15	0.2	0.9	0.3	0.2	0.8 / 0.2 / 0.2

Parameter values used to replicate studies from Williams and Williams [WW69] and Sanabria et al. [San+06], with their interpretation: goal-trackers (GTs) or sign-trackers (STs). * Note that one pigeon of Williams and Williams (P19) behaved as those of Sanabria et al. (i.e. it would be simulated with GTs parameters).

Results

We applied the present model to the various MDPs to replicate the results of Experiments 1, 3 and 4 of Williams and Williams [WW69] and also to some results of Sanabria et al. [San+06] (*Brief PA* protocol).

Classical negative automaintenance

The central phenomenon that we intend to replicate with the present computational model is the greater or lesser persistence in pigeons to peck a key light that, while predictive of reward delivery, leads to its omission in case of contact.

In the first experiment of Williams and Williams [WW69], pigeons undergoing a negative automaintenance procedure failed to completely stop pecking at the key light such that they missed a consequent number of rewards. Only one pigeon (P19) retrieved more than 90% of the available rewards. The model can replicate the general behaviour of all other pigeons with one set of parameter values, and P19 with a different set of values. The red curve in Figure 7.5 shows pigeons that are unable to refrain from pecking and lose almost half of the 50 possible available rewards per session. This behaviour persists over time.

In a more recent study, Sanabria et al. [San+06] challenged these results of Williams and Williams [WW69] as they ran a similar experiment but observed a significant decrease in the detrimental pecks at key light (similar to P19, which was assimilated to a pigeon of Sanabria et al. [San+06] in this experiment). They claimed that remaining pecks did not differ significantly from those that can be observed after a classical extinction procedure. Actually, in an extinction procedure, the conditioned key light is subsequently decorrelated from food delivery, which results in pigeons stopping to emit conditioned responses, except from few exploration pecks. The model is also able to replicate such results using the same MDP despite a slight difference in the experimental protocols. The blue curve in Figure 7.5 shows pigeons that start to peck (by exploration or familiarization) but quickly learn to refrain from pecking to retrieve rewards. We would consider P19 as part of such pigeons.

Each time a simulated pigeon does not peck the key light, its motivational value is reinforced as the key light is contingent to reward delivery (Figure 7.2). This naturally increases the tendency, promoted by the FMF system, to peck during subsequent trials. As in Lesaint et al. [Les+14b], we assume that the presence of the magazine during ITI makes it lose parts of its acquired motivational values (A low u_{ITI}), hence the magazine remains less attractive than the key light and the pigeon never really focuses on it while key light is active. The relative attractiveness of the key light is however balanced by pecks, as the omission of rewards produces a decrease in the key light motivational value.

The MB system solves the task by finding the shortest sequence of actions until reward. As a result, it favours approaches to the magazine, as this is the shortest path to reward (Figure 7.2). Note that other paths would only delay reward delivery by one step and hence are still positively evaluated (especially with a high γ). When close to the key light, it strongly favours refraining from pecking, as this would prevent delivery of the

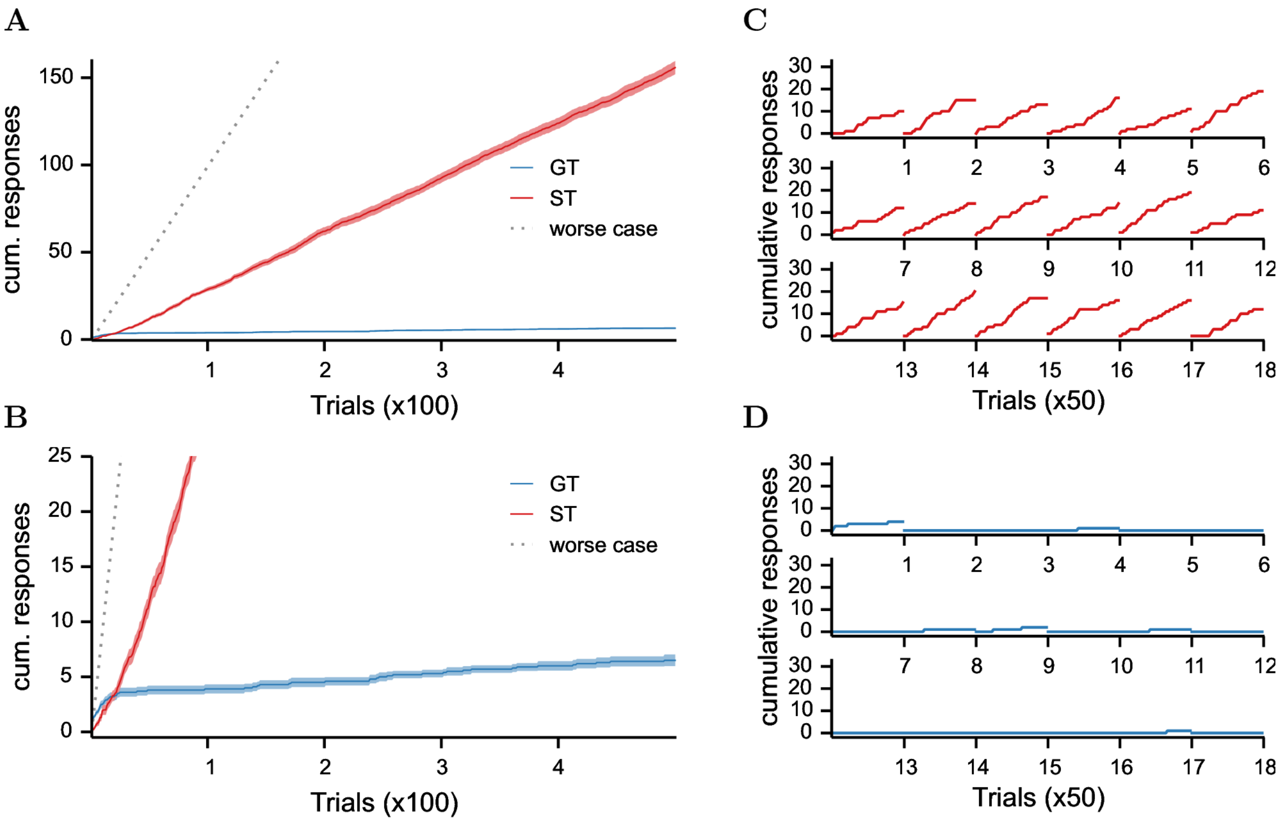


Figure 7.5: Simulation of Experiment 1 of Williams and Williams [WW69] and Brief PA protocol of Sanabria et al. [San+06]. **(A)** *Cumulative pecks towards negative key light made by 8 simulated GT pigeons (blue curve) and 8 simulated ST pigeons (red curve). The dotted grey curve simulated the worse case scenario (if pigeons would have pecked at every trials). Data are expressed as mean ± SEM.* **(B)** *Zoom of (A) for a better reading of the blue curve (GTs).* **(C)** *Cumulative pecks for one ST pigeon by blocks of 50 Trials. To be paralleled with Figure 1 of [WW69].* **(D)** *Cumulative pecks for one GT pigeon by blocks of 50 Trials.*

subsequent reward.

To summarize, in the MB system, the values of all actions but engaging with the key light increase until a convergence level, which depends on how short is the following optimal path to reward. The values then remain at that level until the end of the experiment. The value of engaging the key light remains to 0 as it leads to no reward. In the FMF system, the lever acquires a value that keeps oscillating around a certain level, decreasing at key pecks and increasing otherwise. The magazine value increases at each trial but is partially reset during ITI, such that its value remains at a low level.

When the model gives a high influence (large ω) to the FMF system in the decision process, it produces pigeons that persist in pecking. The FMF system introduces a bias towards actions that lead to approach and interact with stimuli that acquired motivational values, in this case the key light. The resulting low influence of the MB system cannot compensate for this bias. This leads to the production of the expected maladaptive behaviour observed in Williams and Williams pigeons, except for pigeon P19 (Figure 7.5, red curve).

When the model gives a low influence (small ω) to the FMF system in the decision process, it produces pigeons that quickly learn to stop pecking after a few exploration pecks. Indeed, the MB system favours behaviours that maximize cumulation of rewards, that is behaviours that do not lead to peck the key light. Pecks observed in such simulated pigeons are mainly due to exploration. The FMF system is not able to bias the actions enough to lead to a maladaptive behaviour and pigeons stop pecking as in Sanabria et al. [San+06] study and for pigeon P19 of Williams and Williams [WW69] (Figure 7.5, blue curve).

Given the provided equations, refraining from pecking does not completely compensate for a prior peck and vice versa. Combined with exploration, this mechanism leads to oscillations of the behaviour of pigeons that are not a perfect alternation of pecks and abstentions. Hence, from time to time, pigeons will stop pecking, start accumulating food, and by this process reinstate the attractiveness of the key light and the resulting subsequent detrimental pecks.

Thus, the current model is able to account for these, at first sight, contradictory results. With different parameter values (see Table 7.1), the model can reproduce pigeons that fit those of Williams and Williams [WW69] and Sanabria et al. [San+06]. It explains the difference between their findings as a result of a possible interindividual variability in pigeons. Some are more prone to rely on the FMF system to guide their behaviours while others rely on the MB system. We can define the pigeons of Williams and Williams [WW69] as being mainly sign-trackers and those of Sanabria et al. [San+06] as being goal-trackers.

It is important to note that the model describes the significantly lesser amount of reward received by sign-trackers relative to goal-trackers as a consequence and not a cause of their behaviour (simulated by a different ω parameter).

Avoidance strategies

Experiment 2 of Williams and Williams [WW69], using a different protocol, only controlled that key lights had to be contingent to some rewards to produce key pecks and was not simulated. In their Experiments 3 and 4, Williams and Williams [WW69] further investigated the properties of the sustained pecks, especially if they could be oriented to alternative keys with different contingencies (avoidance strategies). A model accounting for negative automaintenance should reproduce these properties.

In Experiment 3, Williams and Williams [WW69] extended the protocol with an additional key light. The new key light would turn on and off at the same time as the previous one, but pecks would have no effect on it, hence named irrelevant key (I). While it seems that pigeons are unable to refrain from pecking, they are still able to orient their pecks towards the less prejudicial target. They observed that in such procedure, a tendency to peck also developed in pigeons, but favouring the irrelevant key, hence maximizing accumulation of rewards. Furthermore, to study if such tendency could be revised once trained, the effect of keys (K and I) was reversed at some point without informing the pigeon, i.e. pecks at the irrelevant key blocked reward delivery and pecks at the negative one were without effect. They observed that pigeons quickly learned to switch to the new irrelevant key (see Figures 5 and 6 of Williams and Williams [WW69]).

With the same parameter values used to simulate Experiment 1 of Williams and Williams [WW69], the model is able to reproduce such properties (Figure 7.6). Simulated pigeons learn to focus on the irrelevant key (I), learn to avoid the negative key (K), and after a unexpected reversal (I becoming negative and K becoming irrelevant), quickly learn to reverse their behaviour.

The irrelevant key provides pigeons with an alternative path, that is more favoured by the model. The rational MB system favours equally well approaches towards the irrelevant and negative keys as there exists a subsequent path of equal length to reach rewards (classical reinforcement learning theory). Hence, the action selected ultimately depends on the bias introduced by the second system. The FMF system gives a higher value to the irrelevant key relative to the negative one, as the irrelevant key is always contingent to reward whereas the negative key is only contingent to reward when no pecks are performed. As a result, orienting towards the irrelevant key has a higher probability of being chosen.

The effect of reversal is better explained through a concrete example. Assuming that the key light K is negative in the current block i, then $\mathcal{V}_i(K) < \mathcal{V}_i(I)$ (V_i denotes the value during block i). When switching to block $i+1$, I becomes irrelevant and $\mathcal{V}(I)$ quickly lowers to the level of $\mathcal{V}_i(K)$ while $\mathcal{V}(K)$ eventually increases to the level of $\mathcal{V}_i(I)$, such that after few trials, $\mathcal{V}_{i+1}(K) > \mathcal{V}_{i+1}(I)$. Hence, the preferred key alternates between each blocks. Hence, the model nicely explains why pigeons cannot refrain from pecking but are still able to orient pecks to a less detrimental key.

In Experiment 4, Williams and Williams [WW69] extended the protocol with another additional key light. The new key light would never turn off and pecks would have no effect on it, hence labelled continuous key. Note that while always lit on, the position of the key (left/right/middle of the key lights panel) was switched after each trial, such that contrary to the fixed magazine, shifts in its positions were predictive of a new possible reward. They studied the relative power of the three keys to attract pecks by combining a subset of them and activating them at different times in different protocols (see Table 7.2).

They observed that all keys, presented alone produced sustained pecks. The continuous key was ineffective in attracting key pecks when an alternative key, either negative (Figure 7 A and C in Williams and Williams [WW69]) or irrelevant (Figure 7 B and D in Williams and Williams [WW69]) was presented. As in Experiment 3, the irrelevant key was effective in attracting away pecks from the negative key (Figure 7 B and D in Williams and Williams [WW69]).

The model is also able to explain these additional results (Figure 7.7). The effectiveness of the irrelevant key to attract key pecks has already been explained for Experiment 3. The ineffectiveness of the continuous key results from its presence during ITI. We hypothesize that the presence of a stimulus within the ITI leads to a decrease of its motivational value. Hence, the motivational value of such a stimulus is lower than those of the alternative keys that are time-locked to reward delivery. Note that for the continuous key to be the

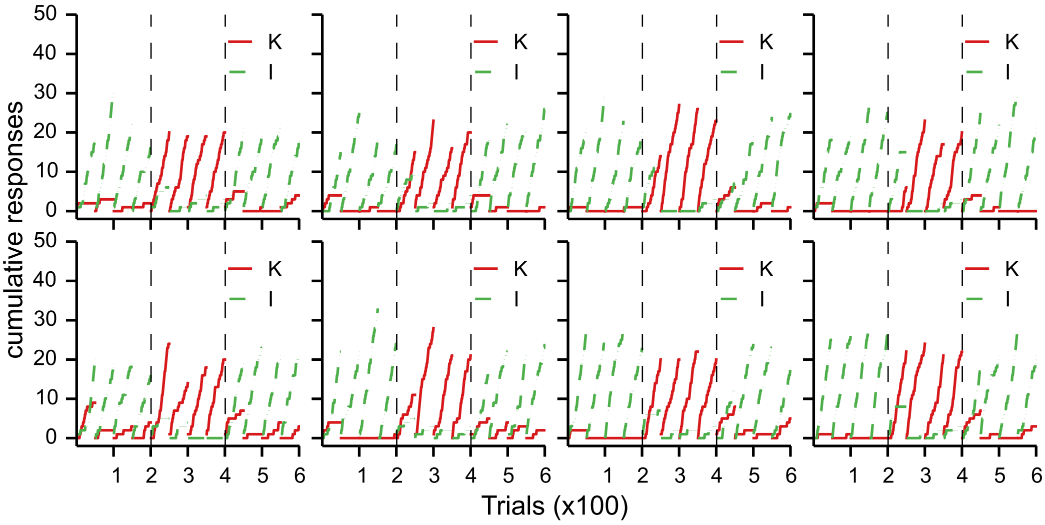


Figure 7.6: Simulation of Experiment 3 of Williams and Williams [WW69]. (A) *Cumulative pecks towards negative key (red curve) and irrelevant key (green curve) over time made by 8 simulated pigeons. Vertical bar indicates reversals of effects between key lights. The dotted grey curve simulated the worse case scenario (if pigeons would have pecked the negative key at every trials). Data are expressed as mean ± SEM.* **(B)** *Cumulative pecks for one pigeon by blocks of 50 Trials. To be paralleled with Figures 5 and 6 of [WW69].*

Table 7.2: Experimental setups for Experiment 4

Protocol	Phase 1	Phase 2	Phase 3
A	K	K + C	C
B	K	K + C + I	C
C	K + C	K + C	C
D	K + C + I	K + C + I	C

Lists of keys activated during the different phases of protocols used in Experiment 4 of Williams and Williams [WW69]. K stands for the negative key, I for the (intermittent) irrelevant key and C for the continuous (irrelevant) key.

focus of pecks when presented alone, its motivational value should however remain higher than the value of the magazine. We do not use the same parameter value to decrease the value of the magazine and the value of the continuous key. A variability in the last parameter could explain why in the experimental data, some pigeons did not engage with this continuous key even presented alone.

Discussion

We applied the model of Lesaint et al. [Les+14b] to a new set of experimental data on a negative automaintenance procedure and showed that it is able to qualitatively reproduce

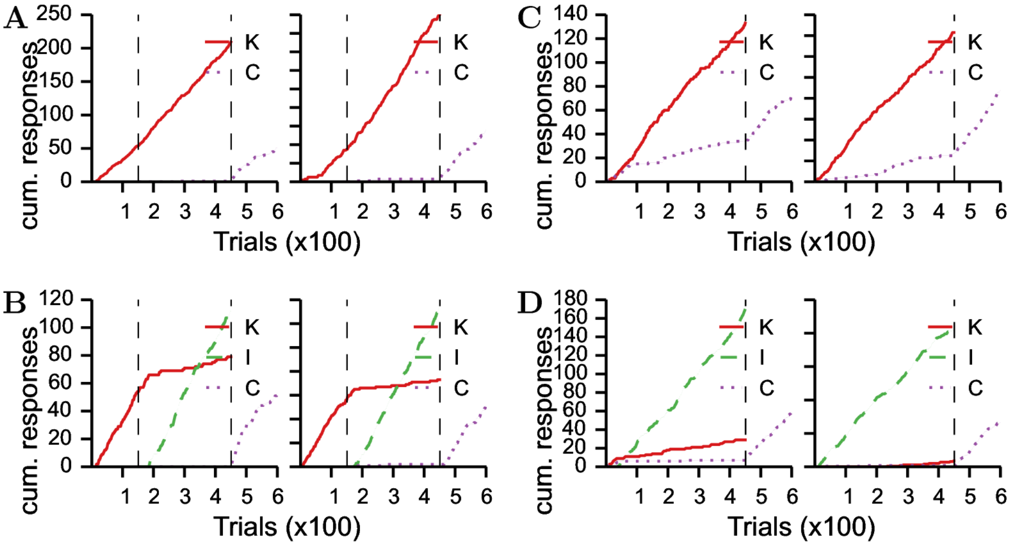


Figure 7.7: Simulation of Experiment 4 of Williams and Williams [WW69]. *Cumulative pecks towards negative key (solid line), irrelevant key (dashed line) and continuous key (dotted line) over time made by 2 simulated pigeons in different protocols (described in Table 7.2). Vertical bar indicates phase switches. To be paralleled with Figure 7 of Williams and Williams [WW69].*

different properties of the resulting phenomenon. This model also provides a plausible explanation, although maybe partial, for the conflictual observations between the studies of Williams and Williams [WW69] and Sanabria et al. [San+06]. It suggests that negative automaintenance arises from the competition of two reinforcement learning systems, one of which relies on factored representations to use values over features rather than states.

Pavlovian and instrumental interactions

In [Les+14b], the computational model was used to account for a phenomenon described as only Pavlovian, hence one could see both systems as different mechanisms of Pavlovian conditioning [DB14]. Here, the same model is used to account for a Pavlovian and instrumental interaction phenomenon and systems are rather seen as each accounting for a different type of conditioning [DB02; Day+06]. Hence, while using a similar Model-Based system for both studies, it might actually reflect different systems in the brain which would rely on similar principles. It is actually unclear if the whole behaviour of rats undergoing autoshaping, from approach to consumption-like engagement, should be classified as purely Pavlovian [Nic10; Huy+11; Geu+13]. Further experiments (e.g. outcome devaluation) should be conducted to clarify this point. Extending from studies on how Pavlovian conditioning affects instrumental tasks [DB02; Yin+08] and studies on how instrumental conditioning can also subsequently affect Pavlovian tasks [AE81; Pré+13], we suggest that many conditioning tasks might present both Pavlovian and instrumental aspects, with one possibly masking the sparse presence of the other.

In the present case, a parallel can be made between Pavlovian conditioning versus instrumental conditioning and the FMF system versus the MB system. Pecks towards key lights arise because of the values they acquire within the FMF system. These motivational values developed solely by contingencies of key lights with food delivery, independently of actions taken. Hence, the FMF system is at the heart of the Pavlovian aspect in simulated pigeons. It biases their actions towards attractive and predictive stimuli, possibly leading to impulsive, and possibly detrimental engagements. Refraining from pecking, on the other side, is learned by the MB system as the appropriate action to get rewarded. Hence, animals know how to act to optimize their rewards. Therefore, the MB system is at the heart of the instrumental aspect of the behaviour of pigeons. It allows them to learn, to some extent, that they must refrain from acting to retrieve food in specific situations, in this case from pecking. We do not state that instrumental conditioning is Model-Based nor Pavlovian conditioning is Model-Free. It has been shown that both aspects are present in both type of conditioning [DB14; Yin+08; BO09]. In the present work, only the Model-Based aspect of instrumental conditioning and the Model-Free aspect of Pavlovian conditioning are sufficient to replicate the data.

The computational model explains the behaviour of pigeons as a combination of both systems. Each system provides valuation informations regarding the current situation, which are further integrated to eventually determine the action to be taken. Moreover, informations are not weighted equally but through a pigeon specific weight (ω) such that one system can have to assess a situation as very detrimental to compensate for the weak positive valuation of this situation attributed by the other system, and avoid a maladaptive behaviour. This is exactly what happens in the negative automaintenance procedure, as the Pavlovian system records the key light as strongly motivational, whereas the instrumental system records any engagement as detrimental. Furthermore, the procedure is such that applying the strategy favoured by one system subsequently reinforce the strategy favoured by the other one. As a result, no system can forever be dominant.

While we currently modelled our integration of MB and FMF systems with a fixed ω parameter, it might be possible, as suggested in the work of Dayan et al. [Day+06] that such weighting parameter would fluctuate over time based on some yet unknown and still debated criterion [Daw+05; Ker+11; Pez+13]. However, we would still expect that subgroups of individuals would show different parameter values and/or that such values would fluctuate differently. The currently investigated data on pigeons cannot rule out an alternative interpretation that, based on a dynamically computed score (e.g. the difference of estimated uncertainty of each system [Daw+05]), only one system might be active and guide the behaviour at a time. However, based on the data about rats undergoing autoshaping experiments simulated with the same model [Les+14b], the full spectrum of observed behaviours ranging from STs to GTs [Mey+12] and the consumption-like engagement of both STs and GTs, explained by the permanently active FMF system, argues against it.

Interestingly, the current model does not necessarily imply that the two systems would favour conflicting policies. For example, in the case of autoshaping [Les+14b] no rewards are lost while the policies favoured are different. Furthermore, the system could even lead to a fruitful collaboration if both systems would favour the same actions, possibly increasing the rate at which the animal would engage with some object and be rewarded accordingly (e.g. in general Pavlovian-to-Instrumental Transfer procedure [CB05; Huy+11; GM+12]). We assume that these systems developed for collaboration rather than competition, as negative automaintenance is not really common in a natural environment. One system provides a rational plan of actions while the other offers the opportunity to accelerate it (e.g. reacting at the shadow of a prey rather than waiting for the prey to be entirely visible). Further investigations will be required to determine whether the collaboration between these systems better explains a variety of animal conditioning behaviours than competition.

Factored representations

Taking advantage of features that compose the environment is not new in the study of Pavlovian conditioning [Sch+96; Bal99; Red+07; SM07; Cou+06; GN12]. It is indeed central to account for phenomena when conflicts arise from the presence of multiple stimuli (e.g. blocking [Kam67] or overexpectation [LN98]). However, the computational models accounting for Pavlovian conditioning phenomena are usually not relying on the classical RL framework (e.g. MDPs or temporal discounting). Furthermore, they mainly tend to describe the varying intensity of a unique conditioned response rather than the variations of observed responses and they do not explain how an agent can learn sequences of actions.

In traditional studies of instrumental tasks, working at the state level is sufficient to reproduce and explain behavioural data [Daw+05; Ker+11; Doy+02; Day+06]. Tasks are defined as standard MDPs, and classical algorithms cannot use the underlying structure to generalize updates to states that share similarities. These models are mainly used to study learning phases and adaptive capabilities in a changing environment, when animals behave near optimally. Classical algorithms are proven to converge to the optimal solution [SB98]. In the current task, without relying on very distinct sets of possibly unusual parameter values, two classical algorithms combined in a model would eventually reach the same optimal policy and hence would fail to explain the variability of observed maladaptive behaviours [Les+14b].

Here factored representations used in one of the two simulated systems but not the other enable these systems to propose different complementary decisions and thus to explain the variety of behaviours observed in the data. Such factored representations are already present in the RL literature and mainly used to overcome the curse of dimensionality [Bel57], i.e. standard algorithms do not scale well to high dimensional spaces and require too much physical space or computation time. Value function approximations [Doy+02; Kha+06; Elf+13] or factored reinforcement learning [Bou+00; Deg+06; VB08] help to build a compact value-function or infer the value of states from values of features. These algorithms are only meant to optimize computations but should not produce outputs that diverge from traditional flat RL algorithms. Here, we use factored representation in a different way and make values over features compete in the choice for the next action. The FMF algorithm generates an output different from traditional RL systems.

The capacity of the model to replicate the maladaptive behaviour of pigeons under negative automaintenance results from the difference between the policies developed by the MB system and the FMF system. Such difference is due to the way factored representations are used by the latter system. While the MB system associates value to general situations (states) and favours an optimal policy, the FMF system associates value to salient stimuli (features) biasing actions towards them and favours a different sub-optimal policy (w.r.t the MDP). The FMF system develops an impetus towards triggering low-level ingrained Pavlovian behaviours towards these salient stimuli as soon as they are presented within a context associated with reward value [Day+06]. In other words, the FMF system and the MB system use different heuristics (paying attention to the situation versus paying attention to salient elements) to guide behaviour. Once combined, these systems conflict in the current experimental setup leading to the observed maladaptive behaviour.

It might be possible to use a factored implementation of the MB system. In such case, we would assume that this system would still assess situations rather than stimuli individually. Hence, it would use factored representations in a traditional way, for computational optimization purposes that should not change the resulting output of the system.

The capacity to attribute values to features also provides a straightforward explanation for why the irrelevant key light attracts most of the pecks in the presence of the negative key light and/or the continuous key light, and why the negative key light attracts most of the pecks in the presence of the continuous key light. Having values over key lights allows for a direct comparison, the development of a preference towards the most valued one, and after its removal, a quick shift towards the second most valued one. By using factored representations to attribute values to features in the classical RL framework, we therefore reunite concepts of the Pavlovian conditioning and instrumental conditioning literature that are rarely combined together, to model some Pavlovian-instrumental interactions.

One must note that the model of Dayan et al. [Day+06] is also able to replicate the results of the first experiments. It also uses a weighted sum between a classical RL system and some impetus system, and by varying the weight of the two systems, it can also produce behaviours that may be paralleled to sign-tracking and goal-tracking. However, in its current form, their model is unable to reproduce the other experiments of Williams and Williams [WW69]. Their impetus system is designed to arbitrary bias the model towards an action a priori defined as Pavlovian, in this case *Go* against *NoGo*, by adding the mean reward value of the ongoing experiment. Introducing new alternative *Go* actions raises questions on whether they should be defined as Pavlovian or not, and on the way they should be biased, i.e. using the same mean reward value or a different one. Even so, it seems that this would not explain the preference for intermittent keys versus continuous keys. While there might be ways to make it work, we think that the use of factored representations makes it straightforward and automatic for our model to explain these experimental data and potentially predict how the model would behave in the presence of new stimuli without filling it with a priori informations. The recording of consumption-like engagements towards the magazine during goal-tracking like behaviours would argue in favour of our model, which predicts the acquisition of some motivational value towards the magazine, whereas the model of Dayan et al. [Day+06] does not.

Resolution of conflicting results

The difference between all pigeons in Williams and Williams [WW69] but P19 and Sanabria et al. [San+06] parallels well with the inter-variability observed by Flagel et al. [Fla+11b] within rats undergoing an autoshaping procedure. In this study, a unique population of rats provided very distinct subgroups. Sign-trackers were prone to engage with the predictive conditioned stimulus (a lever), and goal-trackers were prone to engage with the magazine where food would be delivered as soon as the lever appears. The computational model reproduces the variability of behaviours in pigeons in these two studies in a similar way, based on the varying influence attributed to each system. The simulated pigeons of Sanabria et al. [San+06] mainly rely on the MB system, while those of Williams and Williams [WW69] mainly rely on the FMF system (except for P19). Given the small size of the populations of pigeons involved, one could hope that with a bigger population we could observe within the same study a larger variation of behaviours similar to those of sign-trackers and goal-trackers. Furthermore, it has been shown that populations of rats taken from different vendors (or even different colonies of the same vendor) can show significant differences in their proportion of sign-trackers and goal-trackers [Fit+13]. If confirmed in pigeons, such a result could strengthen our hypothesis. This does not discard that part of the difference in the observed behaviours also comes from the difference in protocols between the two studies.

It is interesting to note that in a study about guinea pigs [PP78], the averaged individual engaged with the conditioned cue under autoshaping phases and switched to engage with the magazine during negative automaintenance phases. Hence, while not engaging with the cue when detrimental, animals could redirect their engagement impulses towards the magazine, in a manner similar to goal-trackers [Fla+11b]. Such a behaviour could easily be explained by the model with the appropriate parameters, i.e. a reasonably high ω with a low u_{ITI}. It would be interesting to know if pigeons in which negative automaintenance is effective would do the same, i.e. whether they would redirect their pecks towards the magazine, if made possible (e.g. no blocking door).

Gamzu and Schwam [GS74] studied negative automaintenance in 4 squirrel monkeys and showed that only one did express a persistent detrimental engagement, and only during early negative automaintenance sessions. They concluded that the procedure fails to produce maladaptive behaviour in these monkeys. Interestingly, the authors state that while key pressing is virtually eliminated, monkeys orient towards the key and occasionally approach it without contact. The model would be able to account for such behaviour with the motivational value of the key sufficiently high to favour approaches towards it rather than the magazine but not high enough so that it would be impossible to refrain from engaging with it. Gamzu and Schwam [GS74] discuss the fact that, contrary to pigeons, the action of key pressing in monkeys is very different from their consumption behaviour, which could be one of the reason of the failure of the negative automaintenance procedure [Mey+14]. Another interpretation, based on the present model, would be that the 4 monkeys are mainly goal-trackers. It might be aslo the case that monkeys and human brains offer a higher level of control in the integration of the two systems.

Predictions

One of the motivations behind the development of the computational model of Lesaint et al. [Les+14b] was to provide an explanation for the particular patterns of DA recordings observed in rats undergoing an autoshaping procedure [Fla+11b], which challenged the classical reward prediction error hypothesis [Sch98; Fio+03]. Assuming that some of the dopaminergic pathways in pigeons share a similar role to those of rats [Gar+05], the computational model gives predictions about what could be expected from physiological recordings in a negative automaintenance procedure (Figure 7.8).

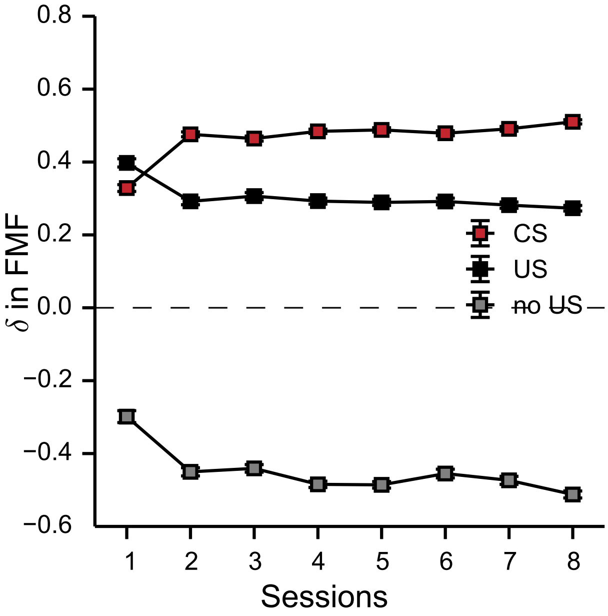


Figure 7.8: Prediction of the model about expected patterns of dopaminergic activity in negative automaintenance. *Data are expressed as mean ± SEM. Average RPE computed by the FMF system at CS appearance (red) and removal of the CS after engagement with the negative key light (no US; gray) and withholding (US; black) for each session of conditioning in the whole population of pigeons (STs and GTs).*

The model predicts that in trials where pigeons orient towards the negative key light (STs or GTs confounded) one should observe DA peaks at CS presentation (as classically expected in such experiments [Sch98]). If pigeons refrain from pecking, one should also observe DA peaks at reward delivery, but with a smaller amplitude (i.e. not a full propagation of DA peaks from the US to the CS as would be expected in an autoshaping experiment). Finally, if pigeons peck the negative key light, one should observe a dip in DA activity when the key light is turned-off and no reward is delivered as expected by the classical omission of an anticipated reward. Note that the model does not use an asymmetrical representation of RPEs, hence it might be possible that DA recordings at pecks might not exactly fit the current prediction [Niv+05a].

Furthermore, the model heavily relies on the hypothesis that the presence of a stimulus, e.g. continuous key light or magazine, during ITI necessarily reduces its value in the FMF system [Les+14b; Les+14a]. Hence, the model predicts that changing the experimental protocol for the ITI part could have some impact on the observed pecks. Indeed, we expect that removing the magazine during ITI, e.g. by blocking it by a door, might make it more attractive to pigeons during key light presentation and hence reduce their detrimental pecks towards any negative key light.

In addition, given that RPEs of the FMF system parallel DA recordings within the core of the nucleus accumbens in rats, we can hypothesize the results of possible lesions or inactivation of the homologue of the dopaminergic system in pigeons. We expect that disabling the FMF system would block any consumption-like behaviour, i.e. pecks towards key lights or magazine. We also expect that pigeons that usually favour approach and engagement towards the key lights will shift their behaviour towards a somewhat more erratic one, i.e. engaging the magazine more often than key lights. Finally, the difference of approach and engagement towards negative, irrelevant and continuous key lights should vanish.

Limitations

As evoked in Lesaint et al. [Les+14b], while using factored representations, and making use of the features within particular states, our approach still relies on the discrete time state paradigm of classical RL, where updates are made at regular intervals and assuming no time required for decisions to be taken. This simplification is sufficient to explain the set of data considered here, however it cannot explain the latencies of responses recorded by Williams and Williams [WW69]. It also prevents us from attempting to qualitatively account for other results of Sanabria et al. [San+06], given that time is an important factor of their protocols.

Model-based capacities of rats have been assessed in multiple studies, however such capacities in pigeons remain to be confirmed. Miyata and Fujita [MF08] showed that pigeons are able to plan one to two steps ahead in mazes, which would confirm their ability to store models of tasks, if simple enough. Further experiments should be conducted to confirm the presence of an MB system in pigeons. Note however that, while the presence of an MB system is necessary to account for the pharmacological data of Flagel et al. [Fla+11b], there is no experimental data on negative automaintenance that requires its presence. A classical MF system would have provided similar results, as both algorithms eventually converge to the same values.

The current results rely on parameters that are hand tuned and could benefit from exhaustive raw data. While we are able to reproduce tendencies and to explain which mechanisms of the model are responsible for them, we could benefit from data on which to actually fit the model more closely, for example by individual trial-by-trial analyses [Daw11]. Additionally, as done by Flagel et al. [Fla+11b], a study that combines not only behavioural data but also physiological and pharmacological data could be of great interest in confirming the model, as previously done by Lesaint et al. [Les+14b].

We did not focus on pretraining conditions and the impact they have on the resulting behaviours. The only possibility offered by the model resides in its initialisation. As in most reinforcement learning studies, with sufficient time, the current model should eventually converge towards a solution that is independent of initial conditions, which is definitely in discrepancy with what was observed. Especially, data tend to show that pigeons need some time to consider pecking, as if some kind of threshold needed to be reached beforehand. The model does not model such aspects of the tasks.

Finally, we did not discuss possible anatomical counterparts of the systems in our computational model, as the involved experiments did not imply any lesions or pharmacological manipulations, e.g. injections of antagonists of the dopamine. Therefore, at the current stage, it would be highly speculative to define which regions of the pigeon brain can be paralleled to each system.

Concluding remarks

Here we used an existing computational model to account for different properties of negative automaintenance, a suggested Pavlovian and instrumental interaction phenomenon. This model was initially developed to account for the variability of behaviours observed in autoshaping experiments [Les+14b]. Interestingly, the account of both autoshaping and negative automaintenance phenomena relies on two major concepts of the model: Dual learning systems and the use of factored representations to use values over features. This works adds to an emerging set of studies suggesting the presence and collaboration of multiple RL systems in the brain. It questions the classical paradigm of state representation and suggests that further investigation of factored representations in RL models of Pavlovian and instrumental processes experiments may be useful to explain their interactions.

Chapter 8

Discussion

All models are wrong, but some models are useful. - George E. P. Box

8.1 Contributions synthesis

In the present thesis, we developed a computational model that combines through a weighted sum a Model-Based RL system (MB) and a Feature-Model-Free RL system (FMF). The FMF system is a Model-Free (MF) system revised to use factored representations to learn values over features and make them possibly compete against each other in the determination of the behaviour [Les+14b]. This computational model contributes to a better formalisation and hopefully understanding of yet unaccounted for experimental data on individual differences during autoshaping in rats [Les+14b; Les+14a]. It provides an explanation to experimental data on maladaptive behaviours during negative automaintenance in pigeons [Les+de]. Finally, it also suggests some new directions in the investigation of the RPE hypothesis of dopamine [Sch+97; Sch98; Sch10; Sch13; Gli11; Har+14; BG05] and incentive salience attribution [BR98; Ber07; Ber12; McC+03; Zha+09; MB09; Tin+09; Lov+11; Mey+12].

The proposed model explains inter-individual differences in rats [Fla+09; Fla+11b; SR12; RF09; Mey+12] and pigeons [WW69; San+06] mainly as the result of a difference in how individuals within a population balance the contribution of the MB and FMF systems. Varying the value of a single weighting continuous parameter explains the continuum of observed behaviours ranging from goal-tracking to sign-tracking. Goal-tracking behaviours are optimal in collecting rewards with regard to the MDP representation of the task. Sign-tracking behaviours are more impulsive and stimuli-oriented. Relying on a single set of parameters values for all experiments, the computational model can link differences observed at the physiological and pharmacological levels to those present at the behavioural level.

The collaboration of both systems also explains maladaptive behaviours [WW69]. By learning values over features, the FMF system may bias the behaviour towards suboptimal solutions that focus on high valued stimuli [Les+14b]. This can be observed in ST rats that engage with the lever under an autoshaping procedure [Fla+09]. The FMF system may even lead to detrimental solutions [Les+de]. This can be observed in the pigeons that peck the conditioned negative key light and consequently block reward delivery under a negative automaintenance procedure [WW69].

We explain the unexpected persistence of dopamine peaks at US time in GT rats [Fla+11b] with respect to the RPE hypothesis of dopamine [Sch+97; Sch98; Sch10; Sch13; Gli11; Har+14; BG05], as an RPE being computed over features rather than over states [Les+14b]. As in classical MF RL models, when the computational model favours a behaviour that engage with a stimulus only partially predictive of reward, an RPE signal remains at the time of the reward following the presentation of this stimulus [Les+14a]. The only partially predictive value of the magazine (focused on by goal-trackers) results from an hypothesized downgrading impact of its value due to its continuous presence during ITI.

Finally, we explain the expression of incentive salience [BR98; Ber07; Ber12; McC+03; Zha+09; MB09; Tin+09; Lov+11; Mey+12], i.e. that some cues become wanted for themselves, by the bias introduced by the FMF system towards high valued stimuli. Such high value, seen as motivational, is acquired as soon as they are contingent and time-locked to reward delivery.

8.2 Limits and perspectives

The present work comes with limits and raises more or less general questions, especially relative to other works in the current field. This section is split in different themes for a better readability, but most of the points discussed can be linked to each other.

8.2.1 Reinforcement Learning framework

Classical RL systems, when simulated for enough time (which can sometimes be very long), are mathematically guaranteed to converge to an optimal solution with regard to the MDP [WD92; Sin+00]. The investigated experimental data [WW69; Fla+11b] are a challenge for pure classical RL algorithms, as no algorithm (MB or MF without degenerative parameter values such as a null learning rate) can explain the unnecessary loss of rewards by pigeons or energy-wasting behaviours of rats. Some models take into account possible constraints (e.g. limits in the memory capacity or reactivity necessity) and develop heuristics to drive the collaboration of multiple systems. Such models are able to express maladaptive behaviours during adaptive phases, at initial learning or after a switch in contingencies [Huy+12; Daw+05; Ker+11; Pez+13]. However, their components relying on the same valuation concepts usually lead to near-optimal solutions after convergence, especially when only one optimal solution exists. By relying on an FMF system that may lead to suboptimal solutions with regard to MDPs, our model can account for maladaptive behaviours [Les+14a]. An alternative approach would have been to set different goals to the different systems of the model or bias specific actions [Day+06], but this would raise the question of how to define such goals and actions. It might also be possible that such maladaptive behaviours reside in mechanisms that can fasten learning in some cases but which would make it detrimental in special cases (e.g. [Huy+11; GM+12; GM+14]). More generally, answering this question involves the study of the heuristics put in place by evolution to deal with limited resources and environmental constraints, the use of reinforcement learning processes being only a part of it.

Given its classical use to account for instrumental conditioning, we chose reinforcement learning as the basic concept of our architecture and we extended it with concepts necessary to account for Pavlovian phenomena, i.e. individual stimuli processing. An alternative could have been to start from a model accounting for Pavlovian phenomena (e.g. the Latent Cause Theory [Cou+04; Cou+06; GN12]) and to extend it with concepts, such as actions, seen as necessary for instrumental conditioning. This approach was used by Cartoni et al. [Car+13] to theoretically account for some Pavlovian-instrumental interaction phenomena and its experimental validation is still under process. Studying such interactions actually raises the question of the difference between instrumental actions and Pavlovian reflexes [Bal+08]. While it seems that humans can distinguish between them (e.g. [Shu+80]), it is unclear whether they should be encoded differently and rely on completely separate circuits. Some actions (e.g. salivating or approaching) can clearly result from one's conscious decision in some cases but also be pure reflexes in others [Huy+11; Nic10; Har+13]. In our current model [Les+14b] as in other models accounting for interactions [Day+06; Car+13], we do not distinguish between actions and reflexes. One way to make this distinction in our model would be to interpret actions that would have naturally been chosen by the MB system as instrumental actions, while those which where only selected because of the bias introduced by the FMF as reflexes.

Based on the integration of values computed by both systems, our model relies on the softmax action selection function to extract actions (and reflexes) that should be taken at each step. This simplification of a complex system is extensively used [Daw+06b; Day+06; Glä+10; Hum+12; Daw+05; Ker+11; Huy+12; Doy+02; Red+07], successful, and generally sufficient in RL models of instrumental conditioning. However, the investigated experimental data show some of its limits, especially regarding temporal aspects. Pigeons seem to never engage with key lights before multiple trials have occurred [WW69]. Furthermore, the latency between rats responses and the reward-predicting cue appearance only diminishes with training [Fla+09]. It is often the case, as in the current experiments [Fla+09; WW69] that animals may take time to engage within a behaviour or a direction [Tol38; Tol39]. It has been suggested that they may resolve conflicts and explore alternatives (e.g. when moving their heads back and forth between multiple available paths [Red+08]). Such reaction-time may increase with the number of options [Hic52; VN00] or reduce over training as behaviour becomes more habitual [Hu+06; Wel80; MR59]. However, the softmax function necessarily selects an action at each time step and sees all actions as being equal – there is no action with a specific treatment. This limitation could be addressed by providing mechanisms that can simulate reaction-time, with extra intermediate states [Daw+06c; KNR09], or introducing some vigour/cost trade-off [Niv+07].

On a more general point of view, this raises the question of the distinction between passively waiting — maybe thinking — and actively refraining from engaging [GM+12; GM+14]. While the latter seem to be an action as any other, could the former be some kind of special default action that should be treated differently [Niv+07] ? This is of particular interest when investigating the impact of lesions or drug injections on the general behaviour. We made the assumption that systemic injections of flupentixol immediately disrupt action selection [Hum+12; Leb+10] (Chapter 5), such that all probabilities to select an action become almost equal, i.e. producing mainly an exploratory behaviour, when it could actually lead to reduce the probabilities of all engagement actions. When we simulate the lesion of one system by setting its contributions to 0, the softmax function immediately rebalances all actions probabilities, while one could expect that the animal would rather reduce the rates of impacted actions without necessarily increasing the rates of alternatives [SR12]. It could be of interest to investigate how more realistic models of action selection in the Basal Ganglia [Kha+05; Gur+04; LG14; Lié+10; Leb+06; Che+13; Bal+13] could actually integrate in the current model or RL-based models in general. Such models indeed need to accumulate some evidences before reaching a decision, i.e. reaching a threshold, and offer hypotheses about anatomical counterparts, that may be better suited to replicate the impacts of specific lesions.

8.2.2 Stimuli processing

In the present model, we only investigated the interest of factored representations on the MF system as it was sufficient to replicate the data. As factored representations are richer than classical ones that ignore the underlying structure of the world, it would be rather logical for other processes to also take advantage of it. Our intuition is that while the FMF system learns values on individual features, we would expect from an MB system using factored representations, to use them to build values over global situations but in a more optimised way, allowing for example to share values between similar states. A computational reason for such intuition is that we would think that a system should be optimal with respect to the information it has access to, in the case of an MB system, its internal model of the world. Factored Reinforcement Learning algorithms (FRL) [Bou+00; Deg+06; VB08; SH04; KS02] are developed on such principle, taking advantage of their knowledge of the structure of the world to build compact representations and fasten computations. This also explains why, to our knowledge, no factored version of MF algorithm have been developed so far, as they do not have access to the structure of the world. If we were to replace our classical MB system with a factored version, it would maintain the important property that both systems should converge to different solutions, given that FRL versions of classical algorithms are not supposed to change the learned solutions [Deg+06; Bou+00; VB08]. Hence, we would expect to obtain the same results. However, while converging to the same solutions, such algorithms can actually behave differently during learning, as their generalization capacities might make them consider new situations as highly valued, or consider impossible situations [Deg+06; Koz+09; Koz10]. Some protocols could be designed to check if some generalization capacities of animals actually match those of such FRL algorithms, and validate their use in computational models of conditioning. For example, if one feature of the experiment has been kept constant during training (e.g. a green light always turned on), then such algorithms usually find it irrelevant, hence, if such a variable changes (e.g. switches to red), one would expect that it has no impact on the expressed behaviour. However, if such a change precedes a modification in reward delivery, then it shall be soon integrated as an important part of the task.

A significant part of our results relies on the hypothesis that the presence of a specific stimulus during ITI makes the value it acquired during conditioning diminish [Les+14b; Les+14a; Les+de]. This hypothesis explains why the magazine is not as attractive as a CS-lever [Fla+11b; RF09], or that a continuously turned on key light will not attract pecks when contingent to a key light that is time-locked to reward delivery even if negative [WW69]. Preliminary work (on a incomplete subset of data) suggests that this devaluation may not be fully explained by a possible engagement towards such elements during ITI (e.g. the amplitude of activity peaks seems uncorrelated with the number of contacts made during ITI). We suggested experimental protocols to confirm that the only presence is sufficient [Les+14a]. An alternative view would be that each stimulus might differ in the value that it can acquire [RW72], depending on its form [Mey+14; Hol77; Mey+10], the conditioning context [BB12; Lub73; Kra+91; Bal+80], or other competition properties as some Pavlovian phenomena could suggest (e.g. overexpectation [LN98; Rey61], overshadowing [Mac76; Res99], or blocking [Kam67; KM96]). Furthermore it has been shown that the ITI duration, relative to the inter stimulus interval, impacts the speed of acquisition and maintenance of conditioned responses to a stimulus [BP79; GB81]. Hence, studying how the duration of ITI or what features are available during this period, could help us clarify its role on stimuli processing and on dopaminergic patterns.

8.2.3 Dopamine signal

The RPE hypothesis of dopamine is not limited to Pavlovian conditioning [Sch+97; Sch98; Sch10; Sch13; Gli11; Har+14; BG05]. Patterns matching RPEs that embed information about actions [WD92; SB98] have also been observed in instrumental tasks [Roe+07; Mor+06; Niv+06; Daw+11; Rob+06]. However, which kind of rule is used to compute such RPEs is still an open question [Bel+12b; Bel+12a; NS08]. Some dopaminergic recordings [Roe+07] suggest a Q-Learning like computation, i.e. the next action is chosen after the system has been updated. In contrast, others [Mor+06] suggest a SARSA-like computation, i.e. the next action has already been chosen and is taken into account at update time. In our model, we arbitrarily chose a Q-Learning-like computation, the next focused feature not being already defined at update time. Further investigation should be required to confirm or revise this choice. However, we would propose that the above conflict might also result from the fact that RPEs could be computed over features/stimuli rather than states, and due to a difference in protocols. In Morris et al. [Mor+06], the two competing stimuli are presented at once. Assuming that monkeys have already chosen the action, hence the stimulus they will engage with, we expect, as observed in the experimental data, a SARSA-like pattern. In Roesch et al. [Roe+07], a unique new stimulus (a new odor) is presented to rats when they are free to choose in which rewarding tunnel they will subsequently go. As only a single stimulus is available and therefore without competitors, even with a SARSA-like rule, we would expect, as observed in the experimental data, a unique pattern of activity whatever the subsequent choice of the rat might be. One could try to redesign the protocol of these studies to use the alternative protocol proposed in the contradictory study and see if the dopaminergic patterns are reversed.

Finally, one must note that the FMF system behaves as an MF system in certain cases (e.g. for STs undergoing an autoshaping procedure [Fla+11b]) and produces RPE patterns that fit expectations from the classical literature [Sch+97; Sch98; Sut88]. Hence, we think that in simple tasks, where there is no stimuli competition such that states and stimuli are equivalent (e.g. second order conditioning [RR72; HR75]) and sign-tracking behaviour seems to dominate, it may be possible to substitute MF systems by FMF systems in current RL-based accounting models. Moreover, in such cases dopaminergic recording might be averaged over trials, sessions or individuals and might hide radically different patterns [Gal+04; Daw11].

We are aware of the active debate over the exact role of dopamine [Red+99a; SC02; Ber07; SC12; Eve14; Nic10; Sch10; Fio+13; Hir14; Gli11; Ste+13]. While we see our system as consistent with the RPE hypothesis of dopamine, we do not challenge here the theory that phasic dopamine might have multiple roles and contribute to multiple systems. Our current implementation does not fit well with a multiphasic decomposition of its roles [Fio+13; Hir14], but is not incompatible with the idea of contributing to multiple functions at once. For example, it may also help in the acquisition of incentive salience [McC+03; Zha+09], given that the learning of the FMF values, assumed to parallel incentive salience, is learned through RPE signals in the FMF system [Ber07; Ber12]. Furthermore, to model the difference of impact between systemic versus local injections of flupentixol, we made the assumption that tonic dopamine can be decorrelated from phasic dopamine, has a specific role [Niv+07; Niv+05b; Hum+12] and is to be linked with action selection [Hum+12; Leb+10]. However, not all models agree on such a clear separation [Fra05; McC+03; Daw+02; Fio+14; She+11]. The exact role and impact of phasic and dopamine activity on learning versus action selection remains currently unclear.

This makes our simulation of the impact of flupentixol rather weak given our limited knowledge on existing experimental data about the precise mechanisms underlying its impact. There are not much evidences on the correct way to model it for both phasic and tonic activity [Hum+12; Leb+10]. We modelled the impact of flupentixol over phasic dopamine in a subtractive way, removing a fixed value of any RPE computed in the model (but see [Red04]). We argued against a multiplicative way, that is by multiplying any RPEs by some small, possibly null, constant, because flattening values would only have slowed learning and not disrupted it as in the experimental data [Fla+11b]. In both cases, such simulations completely neglect the complex dynamics involved and suggest an immediate and fixed effect, which is particularly problematic when the effect of drugs might vary depending on the infusion time (e.g. [SR12]). How to better model the impact of flupentixol, and dopamine antagonists or drugs such as cocaine remains an open question (e.g. see [Pan+07; Red04]).

Finally, experimental data in Roesch et al. [Roe+07] suggest that the expression of conditioned responses is faster than the propagation of dopamine from US time to the decision point, i.e. the informative cue appearance [Bel+12b; Bel+13]. The output of our computational model is only partially based on the FMF system, and it also relies on an MB system assumed to be faster at learning [Daw+05; Ker+11]. We have not yet studied the dynamic of the model nor the investigated experimental data on such aspect, but it would definitely be worth some investigations. It would actually be surprising that a unique signal would explain the whole behaviour, regarding the accumulation of evidences that more than two systems are involved in conditioning [Yin+08; Mee+12; LO12; Mee+10; Mai09].

8.2.4 Pavlovian and instrumental interactions

While the present system is the combination of an MB and an MF system, it is important to distinguish it from models dedicated to instrumental conditioning that also imply these two aspects [Daw+05; Ker+11; Pez+13; Daw+11], or from studies that suggest both aspects in Pavlovian conditioning tasks [Jon+12; RB13]. In the instrumental case, the combination of two systems is used to reproduce the variation in capacities of a unique individual given some criteria (e.g. motivation or training) or experiences [Ott+13; Daw+05; Ker+11; Daw+11; Dol+12; Bei+11]. In the Pavlovian case, it also explains how a unique individual can take into account information from past experiences to immediately revise its behaviour [RB13; Jon+12] in new situations. Our model uses this combination to mainly account for the variability in behaviours of different individuals in the same context, same experience and same conditions. In the present work, we assume that we only modelled the MF aspects of Pavlovian conditioning and MB aspects of instrumental conditioning (GD behaviours). However, some studies suggest that autoshaping is purely Pavlovian [Cla+12; Har+13] and it might be the case that our MB system actually reflect different conditioning aspects depending on the experiments involved [Les+de]. In any case, a more complete model should integrate all aspects (MF/MB instrumental, MF/MB conditioning), and could take inspiration from already existing models. This raises the important question of their integration [Yin+08; Mee+12; LO12; Mee+10; Mai09]. Figure 8.1 illustrates our intuition about this idea. We would suggest that the integration of instrumental and Pavlovian values could actually arise after the competition between the two instrumental systems ($?_1$ in Figure 8.1) [Daw+05; Ker+11] as experiments seem to show that Pavlovian aspects may still bias instrumental actions after overtraining, i.e. when behaviour supposedly became habitual (e.g. in PIT [Hol+10; Wil+12]). We would also suggest that MB aspects of Pavlovian conditioning actually impact MF aspects ($?_2$ in Figure 8.1) before integration [Jon+12; Wie+13;

Tin+09; McD+11; McD+12]. Finally, while we currently modelled our integration of MB and FMF systems with a fixed ω parameter, it might be possible, as suggested in the work of Dayan et al. [Day+06] that it would fluctuate over time based on some yet unknown criterion [Daw+05; Ker+11; Pez+13].

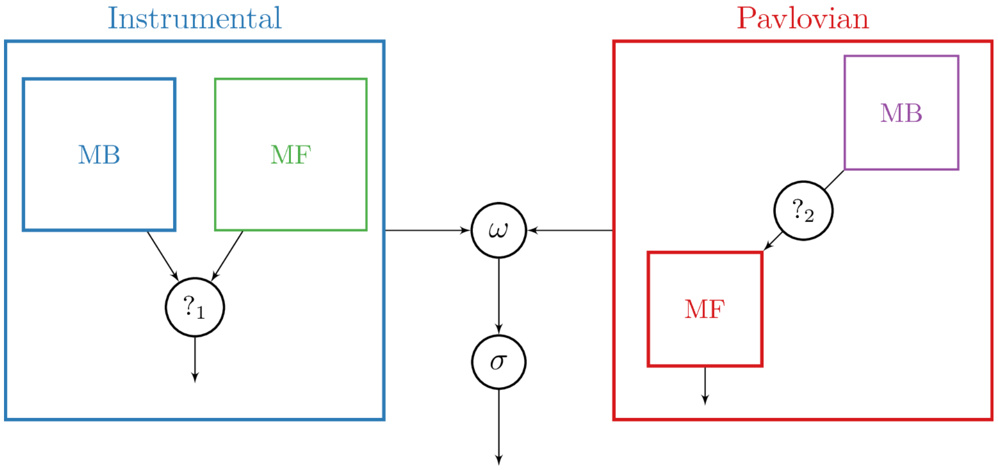


Figure 8.1: Illustration of an hypothesized more complete model. *Instrumental MB and MF subsystems (blue and green inner boxes) compete in the control of the output of the instrumental system (blue outer box). Pavlovian MB subsystem (purple inner box) influences the Pavlovian MF subsystem (red inner box) which provides the output of the Pavlovian system (red outer box). Both outputs are integrated (ω) and subsequently passed to some action selection mechanism (σ).*

We have identified the FMF system as relying, at least partially, on the core of the nucleus accumbens, given the parallel made between its RPE signal and dopaminergic recordings in this region [Fla+11b]. Moreover, simulated behaviours resulting from disruptions of this system were consistent with behaviours of rats with flupentixol injected locally to this region [SR12]. Except from that, we did not investigate much the anatomical counterparts of the other systems in our computational model. Lesions studies, for example of the DMS or OFC, could help identifying which part of the brain is actually accounted for by the MB system in the computational model [Jon+12; Yin+05]. It could also help to clarify whether our assumption that the MB system in the computational model currently accounts for instrumental aspects, rather than some purely Pavlovian aspects, is correct. Flagel et al. [Fla+11a] observed that GTs and STs do not share a similar connectivity of some brain regions, nor the same level of gene expression [Fla+07]. Based on this, we suggest that the current weighted sum integration may result from a crossed projection of brains regions favouring sign-tracking and goal-tracking behaviours (MB and FMF systems) into a third one and that there is a difference in strength of connectivity between such regions in STs vs GTs [Les+14b]. Using techniques such as optogenetics [Ada+11], one could design experiments that manipulate such connectivities and try to make a parallel with our computational model. In the present work, we neglected anatomical differences between species, first in evidence gathering (Chapter 3), second in assuming that rats and pigeons share an MB and an MF systems that would behave similarly, also regarding the potential role of dopamine [Leb+10]. Our work could benefit from a better understanding of differences and similarities between species (e.g. [BO09]), which might help us clarify differences observed on negative automaintenance responses in different species [WW69; GS74; Loc+76; PP78].

Both RL systems of our computational model learn values, over features or states, without keeping information regarding the identity of the reward that was the premise to conditioning. Some phenomena suggest that such identity might not always be taken into account (e.g. [Wil94a; Bur+07]), and that they are based on general incentive properties, such that one can replace a reward with another one without affecting the behaviour. However, other phenomena clearly suggest that identity is an important component (e.g. [McD+11; Jon+12; RB13]). This is especially the case in the PIT phenomenon, where specific PIT occurs only when the currently executed action lead to the same reward than the informative cue played along, while general PIT may enhance any action irrespectively of the reward involved [CB05; CB11; Hol+10; Car+13]. MB systems are often thought as able to embed identity informations, given their internal model, relative to MF systems that would not [Dol+12; McD+11; McD+12; Daw+05].

8.3 Concluding remarks

The present work is a small step towards a unified framework for the study of Pavlovian and instrumental conditioning, especially when experimental tasks might embed both types of conditioning. By taking inspiration from inter-individual differences and maladaptive behaviours observed in conditioning tasks in rats and pigeons, we were able to extract two concepts that seem to be important in such conditioning: multiple reinforcement learning systems and individual stimuli processing and competition. The study of other experimental data might benefit from this approach. We hope that pursuing the investigation of such a combination will provide new insights in the current field.

Publications

Articles

[Les+14c] Florian Lesaint, Olivier Sigaud, Jeremy J Clark, Shelly B Flagel, and Mehdi Khamassi. "Experimental predictions drawn from a computational model of sign-trackers and goal-trackers". In: *J Physiol Paris* (2014). in press.

[Les+14d] Florian Lesaint, Olivier Sigaud, Shelly B Flagel, Terry E Robinson, and Mehdi Khamassi. "Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations". In: *PLoS Comput Biol* 10.2 (2014), e1003466.

[Les+de] Florian Lesaint, Olivier Sigaud, and Mehdi Khamassi. "Accounting for negative automaintenance in pigeons : A dual learning systems approach and factored representations". In: *PLoS One* (under review).

Invited talk

[Les12] Florian Lesaint. "Modelling individual differences observed in Pavlovian autoshaping in rats using a dual - systems view". In: *Inter-GDRs Robotics and Neuroscience Days*. Invited talk. ISIR, UPMC, Paris, 2012.

Posters

[Les+13a] Florian Lesaint, Olivier Sigaud, Shelly B Flagel, Terry E Robinson, and Mehdi Khamassi. "Modelling individual differences in rats using a dual learning systems approach and factored representations". In: *Fifth International Motivational and Cognitive Control Meeting*. Poster. ICM, Paris, 2013.

[Les+13b] Florian Lesaint, Olivier Sigaud, Shelly B Flagel, Terry E Robinson, and Mehdi Khamassi. "Modelling individual differences in rats using a dual learning systems approach and factored representations". In: *First International Interdisciplinary Reinforcement Learning and Decision Making Conference*. Poster. Princeton, USA, 2013.

[Les+13c] Florian Lesaint, Olivier Sigaud, Shelly B Flagel, Terry E Robinson, and Mehdi Khamassi. "Modelling individual differences observed in rats using a dual learning systems approach and factored representations". In: *Third International Interdisciplinary Symposium on Biology of Decision-Making (SBDM 2013)*. Poster. ICM, Paris, 2013.

[Les+14a] Florian Lesaint, Olivier Sigaud, and Mehdi Khamassi. "A model of negative automaintenance in pigeons: dual learning and factored representations". In: *Society for Neuroscience meeting 2014*. Poster. Washington, 2014.

[Les+14b] Florian Lesaint, Olivier Sigaud, and Mehdi Khamassi. "Accounting for negative automaintenance in pigeons: A dual learning systems approach and factored representations". In: *Fourth International Interdisciplinary Symposium on Biology of Decision-Making (SBDM 2014)*. Poster. ICM, Paris, 2014.

Bibliography

[Abr+14] Antony D Abraham, Kim A Neve, and K Matthew Lattal. “Dopamine and extinction: A convergence of theory with fear and reward circuitry”. In: *Neurobiol Learn Mem* 108 (2014), pp. 65–77.

[AD81] Christopher D Adams and Anthony Dickinson. “Instrumental responding following reinforcer devaluation”. In: *Q J Exp Psychol* 33.2 (1981), pp. 109–121.

[Ada+11] Antoine R Adamantidis, Hsing-Chen Tsai, Benjamin Boutrel, Feng Zhang, Garret D Stuber, Evgeny A Budygin, Clara Touriño, Antonello Bonci, Karl Deisseroth, and Luis de Lecea. “Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior”. In: *J Neurosci* 31.30 (2011), pp. 10829–10835.

[Ada82] Christopher D Adams. “Variations in the sensitivity of instrumental responding to reinforcer devaluation”. In: *Q J Exp Psychol* 34.2 (1982), pp. 77–98.

[AE81] Lauren B Alloy and Ronald N Ehrman. “Instrumental to Pavlovian transfer: Learning about response-reinforcer contingencies affects subsequent learning about stimulus-reinforcer contingencies”. In: *Learn Motiv* 12.1 (1981), pp. 109–132.

[AG00] Angelo Arleo and Wulfram Gerstner. “Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity”. In: *Biol Cybern* 83.3 (2000), pp. 287–299.

[Arl+04] Angelo Arleo, Fabrizio Smeraldi, and Wulfram Gerstner. “Cognitive navigation based on nonuniform Gabor space sampling, unsupervised growing networks, and reinforcement learning”. In: *IEEE Trans Neural Netw* 15.3 (2004), pp. 639–652.

[AS12] Eduardo Alonso and Nestor Schmajuk. “Special issue on computational models of classical conditioning guest editors’ introduction”. In: *Learn Behav* 40.3 (2012), pp. 231–240.

[Ash+10] F Gregory Ashby, Benjamin O Turner, and Jon C Horvitz. “Cortical and basal ganglia contributions to habit learning and automaticity”. In: *Trends Cogn Sci* 14.5 (2010), pp. 208–215.

[Bal+07] Bernard W Balleine, Mauricio R Delgado, and Okihide Hikosaka. “The role of the dorsal striatum in reward and decision-making”. In: *J Neurosci* 27.31 (2007), pp. 8161–8165.

[Bal+08] Bernard W Balleine, Nathaniel D Daw, and John P O’Doherty. “Multiple forms of value learning and the function of dopamine”. In: *Neuroeconomics: decision making and the brain* (2008), pp. 367–385.

[Bal+13] Gianluca Baldassarre, Francesco Mannella, Vincenzo G Fiore, Peter Redgrave, Kevin Gurney, and Marco Mirolli. “Intrinsically motivated action–outcome learning and goal-based action recall: a system-level bio-constrained computational model”. In: *Netw* 41 (2013), pp. 168–187.

[Bal+80] Peter D Balsam, CM Locurto, Herbert S Terrace, and John Gibbon. “A search for preexposure effects in autoshaping: Effects of US-only or random CS-UC presentations, intertrial interval duration, and number of pretraining trials.” In: *Psychol Rec* (1980).

[Bal02] Gianluca Baldassarre. “A modular neural-network model of the basal ganglia’s role in learning and selecting motor behaviours”. In: *Cogn Syst Res* 3.1 (2002), pp. 5–13.

[Bal94] Bernard W Balleine. “Asymmetrical interactions between thirst and hunger in Pavlovian-instrumental transfer”. In: *Q J Exp Psychol* 47.2 (1994), pp. 211–231.

[Bal99] Christian Balkenius. “Dynamics of a classical conditioning model”. In: *Auton Robots* 7.1 (1999), pp. 41–56.

[Bar+83] Andrew G Barto, Richard S Sutton, and Charles W Anderson. “Neuronlike adaptive elements that can solve difficult learning control problems”. In: *IEEE Trans Syst Man Cybern* 5 (1983), pp. 834–846.

[Bar95] Andrew G Barto. “Adaptive Critics and the Basal Ganglia”. In: *Models of information processing in the basal ganglia.* Ed. by J C Houk, J L Davis, and D G Beiser. The MIT Press, 1995, pp. 215–232.

[Bau93] William M Baum. “Performances on ratio and interval schedules of reinforcement: Data and theory”. In: *J Exp Anal Behav* 59.2 (1993), pp. 245–264.

[BB00] Jennifer M Birrell and Verity J Brown. “Medial frontal cortex mediates perceptual attentional set shifting in the rat”. In: *J Neurosci* 20.11 (2000), pp. 4320–4324.

[BB08] Stacie L Bancroft and Jason C Bourret. “Generating variable and random schedules of reinforcement using Microsoft Excel macros”. In: *J App Behav Anal* 41.2 (2008), pp. 227–235.

[BB12] Joshua S Beckmann and Michael T Bardo. “Environmental enrichment reduces attribution of incentive salience to a food-associated stimulus”. In: *Behav Brain Sci* 226.1 (2012), pp. 331–334.

[BB61] Keller Breland and Marian Breland. “The misbehavior of organisms.” In: *Am Psychol* 16.11 (1961), p. 681.

[BB79] Mark E Bouton and Robert C Bolles. “Role of conditioned contextual stimuli in reinstatement of extinguished fear.” In: *J Exp Psychol Anim Behav Process* 5.4 (1979), p. 368.

[BD10] Ulrik R Beierholm and Peter Dayan. “Pavlovian-instrumental interaction in ‘observing behavior’”. In: *PLoS Comput Biol* 6.9 (2010), e1000903.

[BD11] Aaron M Bornstein and Nathaniel D Daw. “Multiplicity of control in the basal ganglia: computational roles of striatal subregions”. In: *Curr Opin Neurobiol* 21.3 (2011), pp. 374–380.

[BD12] Aaron M Bornstein and Nathaniel D Daw. “Dissociating hippocampal and striatal contributions to sequential prediction learning”. In: *Eur J Neurosci* 35.7 (2012), pp. 1011–1023.

[BD98a] Bernard W Balleine and Anthony Dickinson. “Goal-directed instrumental action: contingency and incentive learning and their cortical substrates”. In: *Neuropharmacology* 37.4 (1998), pp. 407–419.

[BD98b] Bernard W Balleine and Anthony Dickinson. “The role of incentive learning in instrumental outcome revaluation by sensory-specific satiety”. In: *Anim Learn Behav* 26.1 (1998), pp. 46–59.

[Bei+11] Ulrik R Beierholm, Cedric Anen, Steven Quartz, and Peter Bossaerts. “Separate encoding of model-based and model-free valuations in the human brain”. In: *Neuroimage* 58.3 (2011), pp. 955–962.

[Bel+12a] Jean Bellot, Olivier Sigaud, Matthew R Roesch, Geoffrey Schoenbaum, Benoît Girard, and Mehdi Khamassi. “Dopamine neurons activity in a multi-choice task: reward prediction error or value function?” In: *Proceedings of the French Computational Neuroscience NeuroComp’12 workshop.* 2012, pp. 1–7.

[Bel+12b] Jean Bellot, Olivier Sigaud, and Mehdi Khamassi. “Which Temporal Difference Learning algorithm best reproduces dopamine activity in a multi-choice task?” In: *From Animals to Animats 12.* Springer, 2012, pp. 289–298.

[Bel+13] Jean Bellot, Mehdi Khamassi, Olivier Sigaud, and Benoît Girard. “Which Temporal Difference learning algorithm best reproduces dopamine activity in a multi-choice task?” In: *BMC Neurosci* 14.Suppl 1 (2013), P144.

[Bel57] Richard Bellman. *Dynamic programming.* Princeton University Press, 1957.

[Ber07] Kent C Berridge. “The debate over dopamine’s role in reward: the case for incentive salience”. In: *Psychopharmacology* 191.3 (2007), pp. 391–431.

[Ber12] Kent C Berridge. “From prediction error to incentive salience: mesolimbic computation of reward motivation”. In: *Eur J Neurosci* 35.7 (2012), pp. 1124–1143.

[BG05] Hannah M Bayer and Paul W Glimcher. “Midbrain dopamine neurons encode a quantitative reward prediction error signal”. In: *Neuron* 47.1 (2005), pp. 129–141.

[BJ68] Paul L Brown and Herbert M Jenkins. “Auto-shaping of the Pigeon’s key peck”. In: *J Exp Anal Behav* 11.1 (1968), pp. 1–8.

[BM98] Christian Balkenius and Jan Morén. “Computational models of classical conditioning: a comparative study”. In: *From animals to animats 5: proceedings of the fifth international conference on simulation of adaptive behavior.* MIT Press/Bradford Books: Cambridge, MA. 1998, pp. 348–353.

[BO09] Bernard W Balleine and John P O’Doherty. “Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action”. In: *Neuropsychopharmacology* 35.1 (2009), pp. 48–69.

[Boa77] Robert A Boakes. “Performance on learning to associate a stimulus with positive reinforcement”. In: *Operant-Pavlovian interactions* (1977), pp. 67–97.

[Boo+09] Erie D Boorman, Timothy EJ Behrens, Mark W Woolrich, and Matthew FS Rushworth. “How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action”. In: *Neuron* 62.5 (2009), pp. 733–743.

[Bot+09] Matthew M Botvinick, Yael Niv, and Andrew C Barto. “Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective”. In: *Cognition* 113.3 (2009), pp. 262–280.

[Bot12] Matthew M Botvinick. “Hierarchical reinforcement learning and decision making”. In: *Curr Opin Neurobiol* 22.6 (2012), pp. 956–962.

[Bou+00] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. “Stochastic dynamic programming with factored representations”. In: *Artif Intell* 121.1 (2000), pp. 49–107.

[Bou+12] Mark E Bouton, Neil E Winterbauer, and Travis P Todd. “Relapse processes after the extinction of instrumental learning: Renewal, resurgence, and reacquisition”. In: *Behav Processes* 90.1 (2012), pp. 130–141.

[Bou+14] Mark E Bouton, Travis P Todd, and Samuel P León. “Contextual control of discriminated operant behavior.” In: *J Exp Psychol Anim Learn Cogn* 40.1 (2014), p. 92.

[Bou+95] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. “Exploiting structure in policy construction”. In: *IJCAI*. Vol. 14. 1995, pp. 1104–1113.

[Bou04] Mark E Bouton. “Context and behavioral processes in extinction”. In: *Learn Mem* 11.5 (2004), pp. 485–494.

[BP79] Peter D Balsam and David Payne. “Intertrial interval and unconditioned stimulus durations in autoshaping”. In: *Anim Learn Behav* 7.4 (1979), pp. 477–482.

[BR94] Mark E Bouton and Sean T Ricker. “Renewal of extinguished responding in a second context”. In: *Anim Learn Behav* 22.3 (1994), pp. 317–324.

[BR98] Kent C Berridge and Terry E Robinson. “What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?” In: *Brain Res Rev* 28.3 (1998), pp. 309–369.

[Bro+11] Andrea Brovelli, Bruno Nazarian, Martine Meunier, and Driss Boussaoud. “Differential roles of caudate nucleus and putamen during instrumental learning”. In: *Neuroimage* 57.4 (2011), pp. 1580–1590.

[Bro+12] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. “A survey of monte carlo tree search methods”. In: *Computational Intelligence and AI in Games, IEEE Transactions on* 4.1 (2012), pp. 1–43.

[Bro39] WJ Brogden. “Sensory pre-conditioning.” In: *J Exp Psychol* 25.4 (1939), p. 323.

[BT03] Ronen I Brafman and Moshe Tennenholtz. “R-max-a general polynomial time algorithm for near-optimal reinforcement learning”. In: *The Journal of Machine Learning Research* 3 (2003), pp. 213–231.

[BT14] Mark E Bouton and Travis P Todd. “A fundamental role for context in instrumental learning and extinction”. In: *Behav Processes* 104 (2014), pp. 13–19.

[Bur+07] Kathryn A Burke, Theresa M Franz, Danielle N Miller, and Geoffrey Schoenbaum. “Conditioned reinforcement can be mediated by either outcome-specific or general affective representations”. In: *Front Integr Neurosci* 1 (2007).

[Cal+12] Ken Caluwaerts, Mariacarla Staffa, Steve N’Guyen, Christophe Grand, Laurent Dollé, Antoine Favre-Félix, Benoît Girard, and Mehdi Khamassi. “A biologically inspired meta-control navigation system for the psikharpax rat robot”. In: *Bioinspir Biomim* 7.2 (2012), p. 025009.

[Car+02] Rudolf N Cardinal, John A Parkinson, Jeremy Hall, and Barry J Everitt. “Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex”. In: *Neurosci Biobehav Rev* 26.3 (2002), pp. 321–352.

[Car+13] Emilio Cartoni, Stefano Puglisi-Allegra, and Gianluca Baldassarre. “The three principles of action: a Pavlovian-instrumental transfer hypothesis”. In: *Front Behav Neurosci* 7 (2013).

[CB05] Laura H Corbit and Bernard W Balleine. “Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer”. In: *J Neurosci* 25.4 (2005), pp. 962–970.

[CB11] Laura H Corbit and Bernard W Balleine. “The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell”. In: *J Neurosci* 31.33 (2011), pp. 11786–11794.

[CD83] Gary G Cleland and Graham C L Davey. “Autoshaping in the rat: The effects of localizable visual and auditory signals for food”. In: *J Exp Anal Behav* 40.1 (1983), pp. 47–56.

[CF12] Anne GE Collins and Michael J Frank. “How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis”. In: *Eur J Neurosci* 35.7 (2012), pp. 1024–1035.

[Cha+05] Ricardo Chavarriaga, Thomas Strösslin, Denis Sheynikhovich, and Wulfram Gerstner. “A computational model of parallel navigation systems in rodents”. In: *Neuroinformatics* 3.3 (2005), pp. 223–241.

[Che+13] Fabian Chersi, Marco Mirolli, Giovanni Pezzulo, and Gianluca Baldassarre. “A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning”. In: *Netw* 41 (2013), pp. 212–224.

[CJ07] Laura H Corbit and Patricia H Janak. “Ethanol-Associated Cues Produce General Pavlovian-Instrumental Transfer”. In: *Alcoholism: Clinical and Experimental Research* 31.5 (2007), pp. 766–774.

[CK12] Anne GE Collins and Etienne Koechlin. “Reasoning, learning, and creativity: frontal lobe function and human decision-making”. In: *PLoS Biol* 10.3 (2012), e1001293.

[Cla+12] Jeremy J Clark, Nick G Hollon, and Paul EM Phillips. “Pavlovian valuation systems in learning and decision making”. In: *Curr Opin Neurobiol* 22.6 (2012), pp. 1054–1061.

[Cor+07] Laura H Corbit, Patricia H Janak, and Bernard W Balleine. “General and outcome-specific forms of Pavlovian-instrumental transfer: the effect of shifts in motivational state and inactivation of the ventral tegmental area”. In: *Eur J Neurosci* 26.11 (2007), pp. 3141–3149.

[Cos+13] Ignasi Cos, Mehdi Khamassi, and Benoît Girard. “Modelling the learning of biomechanics and visual planning for decision-making of motor actions”. In: *J Physiol Paris* 107.5 (2013), pp. 399–408.

[Cou+04] Aaron C Courville, Nathaniel D Daw, and David S Touretzky. “Similarity and Discrimination in Classical Conditioning: A Latent Variable Account.” In: *NIPS*. 2004.

[Cou+06] Aaron C Courville, Nathaniel D Daw, and D S Touretzky. “Bayesian theories of conditioning in a changing world”. In: *Trends Cogn Sci* 10.7 (2006), pp. 294–300.

[CR85] Ruth M Colwill and Robert A Rescorla. “Postconditioning devaluation of a reinforcer affects instrumental responding.” In: *J Exp Psychol Anim Behav Process* 11.1 (1985), p. 120.

[CR86] Ruth M Colwill and Robert A Rescorla. “Associative structures in instrumental learning”. In: *Psychol Learn Motiv* 20 (1986), pp. 55–104.

[CS03] Howard C Cromwell and Wolfram Schultz. “Effects of expectations for different reward magnitudes on neuronal activity in primate striatum”. In: *J Neurophysiol* 89.5 (2003), pp. 2823–2838.

[CS11] Paul Chorley and Anil K Seth. “Dopamine-signaled reward predictions generated by competitive excitation and inhibition in a spiking neural network model”. In: *Front Comput Neurosci* 5 (2011).

[Daw+02] Nathaniel D Daw, Sham Kakade, and Peter Dayan. “Opponent interactions between serotonin and dopamine”. In: *Netw* 15.4 (2002), pp. 603–616.

[Daw+03] Nathaniel D Daw, Aaron C Courville, and David S Touretzky. “Timing and partial observability in the dopamine system”. In: *Adv Neural Inf Process Syst* (2003), pp. 99–106.

[Daw+05] Nathaniel D Daw, Yael Niv, and Peter Dayan. “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control”. In: *Nat Neurosci* 8.12 (2005), pp. 1704–1711.

[Daw+06a] Nathaniel D Daw, Yael Niv, and Peter Dayan. “Actions, policies, values and the basal ganglia”. In: *Recent Breakthroughs in Basal Ganglia Research*. Ed. by E Bezard. Nova Science Publishers, Inc Hauppauge, NY, 2006, pp. 91–106.

[Daw+06b] Nathaniel D Daw, John P O’Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. “Cortical substrates for exploratory decisions in humans”. In: *Nature* 441.7095 (2006), pp. 876–879.

[Daw+06c] Nathaniel D Daw, Aaron C Courville, and David S Touretzky. “Representation and timing in theories of the dopamine system”. In: *Neural Comput* 18.7 (2006), pp. 1637–1677.

[Daw+11] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6 (2011), pp. 1204–1215.

[Daw11] Nathaniel D Daw. "Trial-by-trial data analysis using computational models". In: *Decision Making, Affect, and Learning: Attention and Performance XXIII*. Ed. by Mauricio R Delgado, Elizabeth A Phelps, and Trevor W Robbins. Vol. 23. Oxford University Press, 2011. Chap. 1.

[Day+06] Peter Dayan, Yael Niv, Ben Seymour, and Nathaniel D Daw. "The misbehavior of value and the discipline of the will". In: *Neural Netw* 19.8 (2006), pp. 1153–1160.

[Day94] Peter Dayan. "Computational modelling". In: *Curr Opin Neurobiol* 4.2 (1994), pp. 212–217.

[DB02] Peter Dayan and Bernard W Balleine. "Reward, motivation, and reinforcement learning". In: *Neuron* 36.2 (2002), pp. 285–298.

[DB12a] Amir Dezfouli and Bernard W Balleine. "Habits, action sequences and reinforcement learning". In: *Eur J Neurosci* 35.7 (2012), pp. 1036–1051.

[DB12b] Alexandra G DiFeliceantonio and Kent C Berridge. "Which cue to 'want'? Opioid stimulation of central amygdala makes goal-trackers show stronger goal-tracking, just as sign-trackers show stronger sign-tracking". In: *Behav Brain Res* 230.2 (2012), pp. 399–408.

[DB13] Amir Dezfouli and Bernard W Balleine. "Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized". In: *PLoS Comput Biol* 9.12 (2013), e1003364.

[DB14] Peter Dayan and Kent C Berridge. "Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation". In: *Cogn Affect Behav Neurosci* (2014), pp. 1–20.

[DB94] Anthony Dickinson and Bernard W Balleine. "Motivational control of goal-directed action". In: *Anim Learn Behav* 22.1 (1994), pp. 1–18.

[DD13] Ray J Dolan and Peter Dayan. "Goals and Habits in the Brain". In: *Neuron* 80.2 (2013), pp. 312–325.

[DE10] C L Danna and G I Elmer. "Disruption of conditioned reward association by typical and atypical antipsychotics". In: *Pharmacol Biochem Behav* 96.1 (2010), pp. 40–47.

[Dea+99] Richard Dearden, Nir Friedman, and David Andre. "Model based Bayesian exploration". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 150–159.

[Deb+02] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T Meyarivan. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Trans Evol Comput* 6.2 (2002), pp. 182–197.

[Deg+06] Thomas Degris, Olivier Sigaud, and Pierre-Henri Wuillemin. "Learning the structure of factored markov decision processes in reinforcement learning problems". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 257–264.

[Del04] Andrew R Delamater. “Experimental extinction in Pavlovian conditioning: behavioural and neuroscience perspectives”. In: *Q J Exp Psychol B* 57.2 (2004), pp. 97–132.

[Den+01] James C Denniston, Hernán I Savastano, and Ralph R Miller. “The extended comparator hypothesis: Learning by contiguity, responding by relative strength.” In: (2001).

[Dev+99] BD Devan, RJ McDonald, and NM White. “Effects of medial and lateral caudate-putamen lesions on place-and cue-guided behaviors in the water maze: relation to thigmotaxis”. In: *Behav Brain Res* 100.1 (1999), pp. 5–14.

[Dic+95] Anthony Dickinson, Bernard W Balleine, Andrew Watt, Feli Gonzalez, and Robert A Boakes. “Motivational control after extended instrumental training”. In: *Anim Learn Behav* 23.2 (1995), pp. 197–206.

[Diu+13] Carlos Diuk, Anna Schapiro, Natalia Córdova, José Ribas-Fernandes, Yael Niv, and Matthew M Botvinick. “Divide and conquer: hierarchical reinforcement learning and task decomposition in humans”. In: *Computational and Robotic Models of the Hierarchical Organization of Behavior*. Springer, 2013, pp. 271–291.

[DK89] Thomas Dean and Keiji Kanazawa. “A model for reasoning about persistence and causation”. In: *Comput Intell* 5.2 (1989), pp. 142–150.

[DM89] Anthony Dickinson and CW Mulatero. “Reinforcer specificity of the suppression of instrumental performance on a non-contingent schedule”. In: *Behav Processes* 19.1 (1989), pp. 167–180.

[DN08] Peter Dayan and Yael Niv. “Reinforcement learning: the good, the bad and the ugly”. In: *Curr Opin Neurobiol* 18.2 (2008), pp. 185–196.

[Dol+08] Laurent Dollé, Mehdi Khamassi, Benoît Girard, Agnès Guillot, and Ricardo Chavarriaga. “Analyzing interactions between navigation strategies using a computational model of action selection”. In: *Spatial Cognition VI Learning, Reasoning, and Talking about Space*. Springer, 2008, pp. 71–86.

[Dol+10] Laurent Dollé, Denis Sheynikhovich, Benoît Girard, Ricardo Chavarriaga, and Agnès Guillot. “Path planning versus cue responding: a bio-inspired model of switching between navigation strategies”. In: *Biol Cybern* 103.4 (2010), pp. 299–317.

[Dol+12] Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. “The ubiquity of model-based reinforcement learning”. In: *Curr Opin Neurobiol* 22.6 (2012), pp. 1075–1081.

[Dom14] Michael Domjan. *The principles of learning and behavior*. Cengage Learning, 2014.

[Doy+02] Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. “Multiple model-based reinforcement learning”. In: *Neural Comput* 14.6 (2002), pp. 1347–1369.

[Doy00] Kenji Doya. “Reinforcement learning in continuous time and space”. In: *Neural Comput* 12.1 (2000), pp. 219–245.

[Duf95] Steven J Bradtke Michael O Duff. “Reinforcement learning methods for continuous-time Markov decision problems”. In: *Adv Neural Inf Process Syst* 7 (1995), p. 393.

[DW77] James D Deich and Edward A Wasserman. “Rate and temporal pattern of key pecking under autoshaping and omission schedules of reinforcement”. In: *J Exp Anal Behav* 27.2 (1977), pp. 399–405.

[Elf+13] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. “Scaled free-energy based reinforcement learning for robust and efficient learning in high-dimensional state spaces”. In: *Front Neurorobot* 7 (2013).

[Eno+11] Kazuki Enomoto, Naoyuki Matsumoto, Sadamu Nakai, Takemasa Satoh, Tatsuo K Sato, Yasumasa Ueda, Hitoshi Inokawa, Masahiko Haruno, and Minoru Kimura. “Dopamine neurons learn to encode the long-term value of multiple future rewards”. In: *Proceedings of the National Academy of Sciences* 108.37 (2011), pp. 15462–15467.

[Eve14] Barry J Everitt. “Neural and psychological mechanisms underlying compulsive drug seeking habits and drug memories–indications for novel treatments of addiction”. In: *Eur J Neurosci* (2014).

[Fan+13] Rebecca R Fanelli, Jeffrey T Klein, Rebecca M Reese, and Donita L Robinson. “Dorsomedial and dorsolateral striatum exhibit distinct phasic neuronal activity during alcohol self-administration in rats”. In: *Eur J Neurosci* 38.4 (2013), pp. 2637–2648.

[Fau+05] Alexis Faure, Ulrike Haberland, Françoise Condé, and Nicole El Massioui. “Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation”. In: *J Neurosci* 25.11 (2005), pp. 2771–2780.

[Fio+03] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. “Discrete coding of reward probability and uncertainty by dopamine neurons”. In: *Science* 299.5614 (2003), pp. 1898–1902.

[Fio+08] Christopher D Fiorillo, William T Newsome, and Wolfram Schultz. “The temporal precision of reward prediction in dopamine neurons”. In: *Nat Neurosci* 11.8 (2008), pp. 966–973.

[Fio+13] Christopher D Fiorillo, Minryung R Song, and Sora R Yun. “Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli”. In: *J Neurosci* 33.11 (2013), pp. 4710–4725.

[Fio+14] Vincenzo G Fiore, Valerio Sperati, Francesco Mannella, Marco Mirolli, Kevin Gurney, Karl Friston, Raymond J Dolan, and Gianluca Baldassarre. “Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot”. In: *Front Psychol* 5 (2014).

[Fit+13] Christopher J Fitzpatrick, Shyam Gopalakrishnan, Elizabeth S Cogan, Lindsay M Yager, Paul J Meyer, Vedran Lovic, Benjamin T Saunders, Clarissa C Parker, Natalia M Gonzales, and Emmanuel Aryee. “Variation in the Form of Pavlovian Conditioned Approach Behavior among Outbred Male Sprague-Dawley Rats from Different Vendors and Colonies: Sign-Tracking vs. Goal-Tracking”. In: *PLoS One* 8.10 (2013), e75042.

[Fla+07] Shelly B Flagel, Stanley J Watson, Terry E Robinson, and Huda Akil. "Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats". In: *Psychopharmacology* 191.3 (2007), pp. 599–607.

[Fla+09] Shelly B Flagel, Huda Akil, and Terry E Robinson. "Individual differences in the attribution of incentive salience to reward-related cues: Implications for addiction". In: *Neuropharmacology* 56 (2009), pp. 139–148.

[Fla+11a] Shelly B Flagel, Courtney M Cameron, Kristen N Pickup, Stanley J Watson, Huda Akil, and Terry E Robinson. "A food predictive cue must be attributed with incentive salience for it to induce c-fos mRNA expression in cortico-striatal-thalamic brain regions". In: *Neuroscience* 196 (2011), pp. 80–96.

[Fla+11b] Shelly B Flagel, Jeremy J Clark, Terry E Robinson, Leah Mayo, Alayna Czuj, Ingo Willuhn, Christina A Akers, Sarah M Clinton, Paul EM Phillips, and Huda Akil. "A selective role for dopamine in stimulus-reward learning". In: *Nature* 469.7328 (2011), pp. 53–57.

[Foe+06] Karin Foerde, Barbara J Knowlton, and Russell A Poldrack. "Modulation of competing memory systems by distraction". In: *Proceedings of the National Academy of Sciences* 103.31 (2006), pp. 11778–11783.

[Fra05] Michael J Frank. "Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism". In: *J Cogn Neurosci* 17.1 (2005), pp. 51–72.

[Gal+04] Charles R Gallistel, Stephen Fairhurst, and Peter D Balsam. "The learning curve: Implications of a quantitative analysis". In: *Proceedings of the national academy of Sciences of the united States of America* 101.36 (2004), pp. 13124–13131.

[Gar+05] Pascual A Gargiulo, Martin Javier Acerbo, Ines Krug, and JD Delius. "Cognitive effects of dopaminergic and glutamatergic blockade in nucleus accumbens in pigeons". In: *Pharmacol Biochem Behav* 81.4 (2005), pp. 732–739.

[GB81] John Gibbon and Peter D Balsam. "Spreading association in time". In: *Autoshaping and conditioning theory.* Academic Press, 1981, pp. 219–253.

[Ger+13a] Samuel J Gershman, Carolyn E Jones, Kenneth A Norman, Marie-H Monfils, and Yael Niv. "Gradual extinction prevents the return of fear: implications for the discovery of state". In: *Front Behav Neurosci* 7 (2013).

[Ger+13b] Samuel J Gershman, Anna C Schapiro, Almut Hupbach, and Kenneth A Norman. "Neural context reinstatement predicts memory misattribution". In: *J Neurosci* 33.20 (2013), pp. 8590–8595.

[Ger+14] Samuel J Gershman, Arthur B Markman, and A Ross Otto. "Retrospective revaluation in sequential decision making: A tale of two systems." In: *J Exp Psychol Gen* 143.1 (2014), p. 182.

[Geu+13] Dirk EM Geurts, Quentin JM Huys, Hanneke EM den Ouden, and Roshan Cools. "Aversive Pavlovian control of instrumental behavior in humans". In: *J Cogn Neurosci* 25.9 (2013), pp. 1428–1441.

[GG00] Charles R Gallistel and John Gibbon. “Time, rate, and conditioning.” In: *Psychol Rev* 107.2 (2000), pp. 289–344.

[GH72] I Gormezano and George W Hiller. “Omission training of the jaw-movement response of the rabbit to a water US”. In: *Psychon Sci* 29.5 (1972), pp. 276–278.

[Gli11] Paul W Glimcher. “Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis”. In: *Proceedings of the National Academy of Sciences* 108.Supplement 3 (2011), pp. 15647–15654.

[Glä+10] Jan Gläscher, Nathaniel D Daw, Peter Dayan, and John P O’Doherty. “States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning”. In: *Neuron* 66.4 (2010), pp. 585–595.

[GM+12] Marc Guitart-Masip, Quentin JM Huys, Lluis Fuentemilla, Peter Dayan, Emrah Duzel, and Raymond J Dolan. “Go and no-go learning in reward and punishment: interactions between affect and effect”. In: *Neuroimage* 62.1 (2012), pp. 154–166.

[GM+14] Marc Guitart-Masip, Emrah Duzel, Ray J Dolan, and Peter Dayan. “Action versus valence in decision making”. In: *Trends Cogn Sci* 18.4 (2014), pp. 194–202.

[GN12] Samuel J Gershman and Yael Niv. “Exploring a latent cause theory of classical conditioning”. In: *Anim Learn Behav* 40.3 (2012), pp. 255–268.

[GR73] Robert W Griffin and Michael E Rashotte. “A note on the negative automaintenance procedure”. In: *Bull Psychon Soc* 2.6 (1973), pp. 402–404.

[Gra08] Ann M Graybiel. “Habits, rituals, and the evaluative brain”. In: *Annu Rev Neurosci* 31 (2008), pp. 359–387.

[GS74] Elkan Gamzu and Elias Schwam. “Autoshaping and automaintenance of key-press response in squirrel monkeys”. In: *J Exp Anal Behav* 21.2 (1974), pp. 361–371.

[Gur+01] Kevin Gurney, Tony J Prescott, and Peter Redgrave. “A computational model of action selection in the basal ganglia. I. A new functional anatomy”. In: *Biol Cybern* 84.6 (2001), pp. 401–410.

[Gur+04] Kevin. N Gurney, Mark D Humphries, Ric Wood, Tony J Prescott, and Peter Redgrave. “Testing computational hypotheses of brain systems function: a case study with the basal ganglia”. In: *Network* 15.4 (2004), pp. 263–290.

[Hal+01] Jeremy Hall, John A Parkinson, Thomas M Connor, Anthony Dickinson, and Barry J Everitt. “Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour”. In: *Eur J Neurosci* 13.10 (2001), pp. 1984–1992.

[Har+13] Justin A Harris, Benjamin J Andrew, and Dorothy WS Kwok. “Magazine approach during a signal for food depends on Pavlovian, not instrumental, conditioning.” In: *J Exp Psychol Anim Behav Process* 39.2 (2013), p. 107.

[Har+14] Andrew S Hart, Robb B Rutledge, Paul W Glimcher, and Paul EM Phillips. “Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term”. In: *J Neurosci* 34.3 (2014), pp. 698–704.

[Has+10] Mark Haselgrove, Guillem R Esber, John M Pearce, and Peter M Jones. “Two kinds of attention in Pavlovian conditioning: evidence for a hybrid model of learning.” In: *J Exp Psychol Anim Behav Process* 36.4 (2010), p. 456.

[Hau00] Milos Hauskrecht. “Value-function approximations for partially observable Markov decision processes”. In: *J Artif Intell Res* 13.1 (2000), pp. 33–94.

[Her86] Wayne A Hershberger. “An approach through the looking-glass”. In: *Anim Learn Behav* 14.4 (1986), pp. 443–451.

[HG03] Peter C Holland and Michela Gallagher. “Double dissociation of the effects of lesions of basolateral and central amygdala on conditioned stimulus-potentiated feeding and Pavlovian-instrumental transfer”. In: *Eur J Neurosci* 17.8 (2003), pp. 1680–1694.

[Hic52] William E Hick. “On the rate of gain of information”. In: *Q J Exp Psychol (Hove)* 4.1 (1952), pp. 11–26.

[Hir14] Nakahara Hiroyuki. “Multiplexing signals in reinforcement learning with internal models and dopamine”. In: *Curr Opin Neurobiol* 25 (2014), pp. 123–129.

[Hol+10] Nathan M Holmes, Alain R Marchand, and Etienne Coutureau. “Pavlovian to instrumental transfer: a neurobehavioural perspective”. In: *Neurosci Biobehav Rev* 34.8 (2010), pp. 1277–1295.

[Hol+14] Peter C Holland, Judith SA Asem, Connor P Galvin, Caitlin Hepps Keeney, Melanie Hsu, Alexandra Miller, and Vivian Zhou. “Blocking in autoshaped lever-pressing procedures with rats”. In: *Learn Behav* 42.1 (2014), pp. 1–21.

[Hol04] Peter C Holland. “Relations between Pavlovian-instrumental transfer and reinforcer devaluation.” In: *J Exp Psychol Anim Behav Process* 30.2 (2004), p. 104.

[Hol77] Peter C Holland. “Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response.” In: *J Exp Psychol Anim Behav Process* 3.1 (1977), p. 77.

[How60] Ronald A Howard. “Dynamic Programming and Markov Processes”. In: (1960).

[HR75] Peter C Holland and Robert A Rescorla. “Second-order conditioning with food unconditioned stimulus.” In: *J Comp Physiol Psychol* 88.1 (1975), p. 459.

[Hu+06] Dan Hu, Xiaojuan Xu, and Francisco Gonzalez-Lima. “Vicarious trial-and-error behavior and hippocampal cytochrome oxidase activity during Y-maze discrimination learning in the rat”. In: *Int J Neurosci* 116.3 (2006), pp. 265–280.

[Hum+12] Mark D Humphries, Mehdi Khamassi, and Kevin Gurney. “Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia”. In: *Front Neurosci* 6.9 (2012).

[Huy+11] Quentin JM Huys, Roshan Cools, Martin Gölzer, Eva Friedel, Andreas Heinz, Raymond J Dolan, and Peter Dayan. “Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding”. In: *PLoS Comput Biol* 7.4 (2011), e1002028.

[Huy+12] Quentin JM Huys, Neir Eshel, Elizabeth O’Nions, Luke Sheridan, Peter Dayan, and Jonathan P Roiser. “Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees”. In: *PLoS Comput Biol* 8.3 (2012). Ed. by Laurence T Maloney, e1002410.

[Huy+14] Quentin JM Huys, Philippe N Tobler, Gregor Hasler, and Shelly B Flagel. “Chapter 3 - The role of learning-related dopamine signals in addiction vulnerability”. In: *Dopamine*. Ed. by Marco Diana, Gaetano Di Chiara, and Pierfranco Spano. Vol. 211. Elsevier, 2014, pp. 31–77.

[Igl+13] Sandra Iglesias, Christoph Mathys, Kay H Brodersen, Lars Kasper, Marco Piccirelli, Hanneke EM den Ouden, and Klaas E Stephan. “Hierarchical prediction errors in midbrain and basal forebrain during sensory learning”. In: *Neuron* 80.2 (2013), pp. 519–530.

[Jaa+95] Tommi Jaakkola, Satinder P Singh, and Michael I Jordan. “Reinforcement learning algorithm for partially observable Markov decision problems”. In: *Adv Neural Inf Process Syst* 7 (1995), p. 345.

[Jac+91] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. “Adaptive mixtures of local experts”. In: *Neural Comput* 3.1 (1991), pp. 79–87.

[Jam+12] Randall K Jamieson, Matthew JC Crump, and Samuel D Hannah. “An instance theory of associative learning”. In: *Learn Behav* 40.1 (2012), pp. 61–82.

[Jar+06] Elvia Jara, Javier Vila, and Antonio Maldonado. “Second-order conditioning of human causal learning”. In: *Learn Motiv* 37.3 (2006), pp. 230–246.

[Jon+12] Joshua L Jones, Guillem R Esber, Michael A McDannald, Aaron J Gruber, Alex Hernandez, Aaron Mirenzi, and Geoffrey Schoenbaum. “Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values”. en. In: *Science* 338.6109 (Nov. 2012), pp. 953–956. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1227489. URL: http://www.sciencemag.org/content/338/6109/953 (visited on 03/25/2013).

[Kae+96] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. “Reinforcement Learning: A Survey”. In: *J Artif Intell Res* 4 (1996), pp. 237–285.

[Kam67] Leon J Kamin. “Predictability, surprise, attention, and conditioning”. In: *Punishment and aversive behavior*. Ed. by B A Campbell and R M Church. New York: Appleton-Century-Crofts, 1967, pp. 279–296.

[KC03] Simon Killcross and Etienne Coutureau. “Coordination of actions and habits in the medial prefrontal cortex of rats”. In: *Cereb Cortex* 13.4 (2003), pp. 400–408.

[KD10] Yutaka Kosaki and Anthony Dickinson. “Choice and contingency in the development of behavioral autonomy during instrumental conditioning.” In: *J Exp Psychol Anim Behav Process* 36.3 (2010), p. 334.

[Ker+11] Mehdi Keramati, Amir Dezfouli, and Payam Piray. “Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes”. In: *PLoS Comput Biol* 7.5 (2011), e1002055.

[KH12] Mehdi Khamassi and Mark D Humphries. “Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies”. In: *Front Behav Neurosci* 6.79 (2012).

[Kha+05] Mehdi Khamassi, Loïc Lachèze, Benoît Girard, Alain Berthoz, and Agnès Guillot. “Actor–Critic models of reinforcement learning in the basal ganglia: from natural to artificial rats”. In: *Adapt Behav* 13.2 (2005), pp. 131–148.

[Kha+06] Mehdi Khamassi, Louis-Emmanuel Martinet, and Agnès Guillot. “Combining self-organizing maps with mixtures of experts: application to an actor-critic model of reinforcement learning in the basal ganglia”. In: *From Animals to Animats 9*. Springer, 2006, pp. 394–405.

[Kha+08] Mehdi Khamassi, Antonius B Mulder, Eiichi Tabuchi, Vincent Douchamps, and Sidney I Wiener. “Anticipatory reward signals in ventral striatal neurons of behaving rats”. In: *Eur J Neurosci* 28.9 (2008), pp. 1849–1866.

[Kil03] Peter R Killeen. “Complex dynamic processes in sign tracking with an omission contingency (negative automaintenance).” In: *J Exp Psychol Anim Behav Process* 29.1 (2003), p. 49.

[KM96] Yacoub Khallad and Jay Moore. “Blocking, unblocking, and overexpectation in autoshaping with pigeons”. In: *J Exp Anal Behav* 65.3 (1996), pp. 575–591.

[KNR09] Zeb Kurth-Nelson and A David Redish. “Temporal-difference reinforcement learning with distributed representations”. In: *PLoS One* 4.10 (2009), e7362.

[Kon48] Jerzy Konorski. “Conditioned reflexes and neuron organization.” In: (1948).

[Koz+09] Olga Kozlova, Olivier Sigaud, Pierre-Henri Wuillemin, and Christophe Meyer. “Considering unseen states as impossible in factored reinforcement learning”. In: *Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 721–735.

[Koz10] Olga Kozlova. “Apprentissage par renforcement hiérarchique et factorisé”. PhD thesis. Université Pierre et Marie Curie-Paris VI, 2010.

[Kra+91] Philipp J Kraemer, Christopher K Randall, and Timothy J Carbary. “Release from latent inhibition with delayed testing”. In: *Anim Learn Behav* 19.2 (1991), pp. 139–145.

[KS02] Michael Kearns and Satinder Singh. “Near-optimal reinforcement learning in polynomial time”. In: *Mach Learn* 49.2-3 (2002), pp. 209–232.

[KS06] Levente Kocsis and Csaba Szepesvári. “Bandit based monte-carlo planning”. In: *Machine Learning: ECML 2006*. Springer, 2006, pp. 282–293.

[KS08] Shunsuke Kobayashi and Wolfram Schultz. “Influence of reward delays on responses of dopamine neurons”. In: *J Neurosci* 28.31 (2008), pp. 7837–7846.

[KS12] Munir G Kutlu and Nestor A Schmajuk. “Solving Pavlov’s puzzle: Attentional, associative, and flexible configural mechanisms in classical conditioning”. In: *Learn Behav* 40.3 (2012), pp. 269–291.

[Leb+06] Arthur Leblois, Thomas Boraud, Wassilios Meissner, Hagai Bergman, and David Hansel. “Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia”. In: *J Neurosci* 26.13 (2006), pp. 3567–3583.

[Leb+10] Arthur Leblois, Benjamin J Wendel, and David J Perkel. “Striatal dopamine modulates basal ganglia output and regulates social context-dependent behavioral variability through D1 receptors”. In: *J Neurosci* 30.16 (2010), pp. 5730–5743.

[Les+14a] Florian Lesaint, Olivier Sigaud, Jeremy J Clark, Shelly B Flagel, and Mehdi Khamassi. “Experimental predictions drawn from a computational model of sign-trackers and goal-trackers”. In: *J Physiol Paris* (2014). in press.

[Les+14b] Florian Lesaint, Olivier Sigaud, Shelly B Flagel, Terry E Robinson, and Mehdi Khamassi. “Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations”. In: *PLoS Comput Biol* 10.2 (2014), e1003466.

[Les+de] Florian Lesaint, Olivier Sigaud, and Mehdi Khamassi. “Accounting for negative automaintenance in pigeons : A dual learning systems approach and factored representations”. In: *PLoS One* (under review).

[LG14] Jean Liénard and Benoît Girard. “A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection”. In: *J Comput Neurosci* 36.3 (2014), pp. 445–468.

[Lié+10] Jean Liénard, Agnès Guillot, and Benoît Girard. “Multi-objective evolutionary algorithms to investigate neurocomputational issues: the case study of basal ganglia models”. In: *From Animals to Animats 11*. Springer, 2010, pp. 597–606.

[LN98] K Matthew Lattal and Sadahiko Nakajima. “Overexpectation in appetitive Pavlovian and instrumental conditioning”. In: *Anim Learn Behav* 26.3 (1998), pp. 351–360.

[LO12] Mimi Liljeholm and John P O’Doherty. “Contributions of the striatum to learning, motivation, and performance: an associative account”. In: *Trends Cogn Sci* 16.9 (2012), pp. 467–475.

[Loc+76] Charles Locurto, Herbert S Terrace, and John Gibbon. “Autoshaping, random control, and omission training in the rat”. In: *J Exp Anal Behav* 26.3 (1976), pp. 451–462.

[Loc+78] Charles M Locurto, Herbert S Terrace, and John Gibbon. “Omission training (negative automaintenance) in the rat: Effects of trial offset”. In: *Bull Psychon Soc* 12.1 (1978), pp. 11–14.

[Lom+11] Anna M Lomanowska, Vedran Lovic, Michael J Rankine, Skyler J Mooney, Terry E Robinson, and Gary W Kraemer. “Inadequate early social experience increases the incentive salience of reward-related cues in adulthood”. In: *Behav Brain Res* 220.1 (2011), pp. 91–99.

[Lov+11] Vedran Lovic, Benjamin T Saunders, Lindsay M Yager, and Terry E Robinson. “Rats prone to attribute incentive salience to reward cues are also prone to impulsive action”. In: *Behav Brain Res* 223.2 (2011), pp. 255–261.

[Lov83] Peter F Lovibond. “Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus.” In: *J Exp Psychol Anim Behav Process* 9.3 (1983), p. 225.

[LS08] José A Larrauri and Néstor A Schmajuk. “Attentional, associative, and configural mechanisms in extinction.” In: *Psychol Rev* 115.3 (2008), p. 640.

[Lub73] Robert E Lubow. “Latent inhibition.” In: *Psychol Bull* 79.6 (1973), p. 398.

[Lud+12] Elliot A Ludvig, Richard S Sutton, and E James Kehoe. “Evaluating the TD model of classical conditioning”. In: *Learn Behav* 40.3 (2012), pp. 305–319.

[LW08] Hiu Tin Leung and R Frederick Westbrook. “Spontaneous recovery of extinguished fear responses deepens their extinction: a role for error-correction mechanisms.” In: *J Exp Psychol Anim Behav Process* 34.4 (2008), p. 461.

[MA93] Andrew W Moore and Christopher G Atkeson. “Prioritized sweeping: Reinforcement learning with less data and less time”. In: *Mach Learn* 13.1 (1993), pp. 103–130.

[Mac76] N J Mackintosh. “Overshadowing and stimulus intensity”. In: *Anim Learn Behav* 4.2 (1976), pp. 186–192.

[Mai09] Tiago V Maia. “Reinforcement learning, conditioning, and the brain: Successes and challenges”. In: *Cogn Affect Behav Neurosci* 9.4 (2009), pp. 343–364.

[Mar+11] Louis-Emmanuel Martinet, Denis Sheynikhovich, Karim Benchenane, and Angelo Arleo. “Spatial learning and action planning in a prefrontal cortical network model”. In: *PLoS Comput Biol* 7.5 (2011), e1002045.

[Mar+13] Stephen Maren, K Luan Phan, and Israel Liberzon. “The contextual brain: implications for fear conditioning, extinction and psychopathology”. In: *Nat Rev Neurosci* 14.6 (2013), pp. 417–428.

[MB09] Stephen V Mahler and Kent C Berridge. “Which cue to ”want?” Central amygdala opioid activation enhances and focuses incentive salience on a prepotent reward cue.” In: *J Neurosci* 29.20 (2009), pp. 6500–13.

[MC01] Earl K Miller and Jonathan D Cohen. “An integrative theory of prefrontal cortex function”. In: *Annu Rev Neurosci* 24.1 (2001), pp. 167–202.

[McC+03] Samuel M McClure, Nathaniel D Daw, and P Read Montague. “A computational substrate for incentive salience”. In: *Trends Neurosci* 26.8 (2003), pp. 423–428.

[McD+11] Michael A McDannald, Federica Lucantonio, Kathryn A Burke, Yael Niv, and Geoffrey Schoenbaum. “Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning”. In: *J Neurosci* 31.7 (2011), pp. 2700–2705.

[McD+12] Michael A McDannald, Yuji K Takahashi., Nina Lopatina, Brad W Pietras, Josh L Jones, and Geoffrey Schoenbaum. “Model-based learning and the contribution of the orbitofrontal cortex to the model-free world”. In: *Eur J Neurosci* 35.7 (2012), pp. 991–996.

[McD+14] Michael A McDannald, Joshua L Jones, Yuji K Takahashi, and Geoffrey Schoenbaum. “Learning theory: A driving force in understanding orbitofrontal function”. In: *Neurobiol Learn Mem* 108 (2014), pp. 22–27.

[MD10] Jean-Baptiste Mouret and Stéphane Doncieux. “SFERESv2: Evolvin’ in the Multi-Core World”. In: *WCCI 2010 IEEE World Congress on Computational Intelligence, Congress on Evolutionary Computation (CEC).* 2010, pp. 4079–4086.

[Mee+10] Matthijs van der Meer, Adam Johnson, Neil C Schmitzer-Torbert, and A David Redish. “Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task”. In: *Neuron* 67.1 (2010), pp. 25–32.

[Mee+12] Matthijs van der Meer, Zeb Kurth-Nelson, and A David Redish. “Information processing in decision-making systems”. In: *Neuroscientist* 18.4 (2012), pp. 342–359.

[Mey+10] Paul J Meyer, J Wayne Aldridge, and Terry E Robinson. “Auditory and Visual Cues Are Differentially Attributed with Incentive Salience but Similarly Affected by Amphetamine”. In: *Society for Neuroscience Annual Meeting (SfN10)*. 2010.

[Mey+12] Paul J Meyer, Vedran Lovic, Benjamin T Saunders, Lindsay M Yager, Shelly B Flagel, Jonathan D Morrow, and Terry E Robinson. “Quantifying Individual Variation in the Propensity to Attribute Incentive Salience to Reward Cues”. In: *PLoS One* 7.6 (2012), e38987.

[Mey+14] Paul J Meyer, Elizabeth S Cogan, and Terry E Robinson. “The Form of a Conditioned Stimulus Can Influence the Degree to Which It Acquires Incentive Motivational Properties”. In: *PLoS One* 9.6 (2014), e98163.

[MF08] Hiromitsu Miyata and Kazuo Fujita. “Pigeons (Columba livia) plan future moves on computerized maze tasks”. In: *Anim Cogn* 11.3 (2008), pp. 505–516.

[Min96] Jonathon W Mink. “The basal ganglia: focused selection and inhibition of competing motor programs.” In: *Prog Neurobiol* 50.4 (1996), pp. 381–425.

[Mir+13] Marco Mirolli, Vieri G Santucci, and Gianluca Baldassarre. “Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: a simulated robotic study”. In: *Netw* 39 (2013), pp. 40–51.

[MM14a] Bridget L McConnell and Ralph R Miller. “Associative accounts of recovery-from-extinction effects”. In: *Learn Motiv* 46 (2014), pp. 1–15.

[MM14b] Mikael Molet and Ralph R Miller. “Timing: an attribute of associative learning”. In: *Behav Processes* 101 (2014), pp. 4–14.

[MM88] Ralph R Miller and Louis D Matzel. “The comparator hypothesis: A response rule for the expression of associations”. In: *Psychol Learn Motiv* 22 (1988), pp. 51–92.

[Mol+12] Mikael Molet, Gonzalo Miguez, Henry X Cham, and Ralph R Miller. “When does integration of independently acquired temporal relationships take place?” In: *J Exp Psychol Anim Behav Process* 38.4 (2012), p. 369.

[Mor+06] Genela Morris, Alon Nevet, David Arkadir, Eilon Vaadia, and Hagai Bergman. “Midbrain dopamine neurons encode decisions for future action”. In: *Nat Neurosci* 9.8 (2006), pp. 1057–1063.

[MR11] Matthijs van der Meer and A David Redish. “Ventral striatum: a critical look at models of learning and evaluation”. In: *Curr Opin Neurobiol* 21.3 (2011), pp. 387–392.

[MR59] GH Mowbray and MV Rhoades. “On the reduction of choice reaction times with practice”. In: *Q J Exp Psychol (Hove)* 11.1 (1959), pp. 16–23.

[Nak+00] Sadahiko Nakajima, Sayaka Tanaka, Kouji Urushihara, and Hiroshi Imada. “Renewal of extinguished lever-press responses upon return to the training context”. In: *Learn Motiv* 31.4 (2000), pp. 416–431.

[Nak+04] Hiroyuki Nakahara, Hideaki Itoh, Reiko Kawagoe, Yoriko Takikawa, and Okihide Hikosaka. “Dopamine neurons can represent context-dependent prediction error”. In: *Neuron* 41.2 (2004), pp. 269–280.

[Nev+01] John A Nevin, Randolph C Grace, Shasta Holland, and Anthony P McLean. “Variable-ratio versus variable-interval schedules: Response rate, resistance to change, and preference”. In: *J Exp Anal Behav* 76.1 (2001), pp. 43–74.

[Nic10] Saleem M Nicola. “The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior”. In: *J Neurosci* 30.49 (2010), pp. 16585–16600.

[Niv+05a] Yael Niv, Michael O Duff, and Peter Dayan. “Dopamine, uncertainty and TD learning”. In: *Behav Brain Funct* 1.6 (2005), pp. 1–9.

[Niv+05b] Yael Niv, Nathaniel D Daw, and Peter Dayan. “How fast to work: Response vigor, motivation and tonic dopamine”. In: *NIPS*. Vol. 18. 2005, pp. 1019–1026.

[Niv+06] Yael Niv, Nathaniel D Daw, and Peter Dayan. “Choice values”. In: *Nat Neurosci* 9.8 (2006), pp. 987–988.

[Niv+07] Yael Niv, Nathaniel D Daw, Daphna Joel, and Peter Dayan. “Tonic dopamine: opportunity costs and the control of response vigor”. In: *Psychopharmacology* 191.3 (2007), pp. 507–520.

[Niv09] Yael Niv. “Reinforcement learning in the brain”. In: *J Math Psychol* 53.3 (2009), pp. 139–154.

[NS08] Yael Niv and Geoffrey Schoenbaum. “Dialogues on prediction errors”. In: *Trends Cogn Sci* 12.7 (2008), pp. 265–272.

[O’C79] Michael F O’Connell. “Temporal distributions of responding during discrete-trial omission training in rats”. In: *J Exp Anal Behav* 31.1 (1979), p. 31.

[O’D+04] John O’Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. “Dissociable roles of ventral and dorsal striatum in instrumental conditioning”. In: *Science* 304.5669 (2004), pp. 452–454.

[OB05] Sean B Ostlund and Bernard W Balleine. “Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning”. In: *J Neurosci* 25.34 (2005), pp. 7763–7770.

[OB07] Sean B Ostlund and Bernard W Balleine. “Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning”. In: *J Neurosci* 27.18 (2007), pp. 4819–4825.

[Ott+13] A Ross Otto, Candace M Raio, Alice Chiang, Elizabeth A Phelps, and Nathaniel D Daw. “Working-memory capacity protects model-based learning from stress”. In: *Proceedings of the National Academy of Sciences* 110.52 (2013), pp. 20941–20946.

[Pan+07] Leigh V Panlilio, Eric B Thorndike, and Charles W Schindler. “Blocking of conditioning to a cocaine-paired stimulus: testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward”. In: *Pharmacol Biochem Behav* 86.4 (2007), pp. 774–777.

[Par+02] John A Parkinson, JW Dalley, Rudolf N Cardinal, A Bamford, B Fehnert, G Lachenal, N Rudarakanchana, KM Halkerston, TW Robbins, and Barry J Everitt. “Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function”. In: *Behav Brain Sci* 137.1 (2002), pp. 149–163.

[Par+99] John A Parkinson, Mary C Olmstead, Lindsay H Burns, Trevor W Robbins, and Barry J Everitt. “Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive pavlovian approach behavior and the potentiation of conditioned reinforcement and locomotor activity byd-amphetamine”. In: *J Neurosci* 19.6 (1999), pp. 2401–2411.

[Pav27] Ivan P Pavlov. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* Oxford University Press, 1927.

[Pea+98] John M Pearce, Amanda DL Roberts, and Mark Good. “Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors”. In: *Nature* 396.6706 (1998), pp. 75–77.

[Pez+13] Giovanni Pezzulo, Francesco Rigoli, and Fabian Chersi. “The mixed instrumental controller: using value of information to combine habitual choice and mental simulation”. In: *Front Psychol* 4 (2013).

[PM12] Marsha R Penner and Sheri J Y Mizumori. “Neural systems analysis of decision making during goal-directed navigation”. In: *Prog Neurobiol* 96.1 (2012), pp. 96–135.

[Pol+13] Cody W Polack, Mikael Molet, Gonzalo Miguez, and Ralph R Miller. “Associative structure of integrated temporal relationships”. In: *Learn Behav* 41.4 (2013), pp. 443–454.

[PP78] Alan Poling and Teresa Poling. “Automaintenance in Guinea Pigs: Effects of feeding regimen and omission training”. In: *J Exp Anal Behav* 30.1 (1978), pp. 37–46.

[Pré+13] Charlotte Prévost, Daniel McNamee, Ryan K Jessup, Peter Bossaerts, and John P O’Doherty. “Evidence for model-based computations in the human amygdala during Pavlovian conditioning”. In: *PLoS Comput Biol* 9.2 (2013), e1002918.

[Put95] Martin L Puterman. “Markov decision processes: Discrete stochastic dynamic programming”. In: *Journal of the Operational Research Society* 46.6 (1995), pp. 792–792.

[PW93] Jing Peng and Ronald J Williams. “Efficient learning and planning within the Dyna framework”. In: *Adapt Behav* 1.4 (1993), pp. 437–454.

[Rag+02] Michael E Ragozzino, Katharine E Ragozzino, Sheri JY Mizumori, and Raymond P Kesner. “Role of the dorsomedial striatum in behavioral flexibility for response and visual cue discrimination learning.” In: *Behav Neurosci* 116.1 (2002), p. 105.

[RB13] Mike JF Robinson and Kent C Berridge. “Instant Transformation of Learned Repulsion into Motivational “Wanting””. In: *Current Biology* 23.4 (2013), pp. 282–289.

[Red+07] A David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. "Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling." In: *Psychol Rev* 114.3 (2007), pp. 784–805.

[Red+08] A David Redish, Steve Jensen, and Adam Johnson. "A unified framework for addiction: vulnerabilities in the decision process". In: *Behav Brain Sci* 31.04 (2008), pp. 415–437.

[Red+99a] Peter Redgrave, Tony J Prescott, and Kevin Gurney. "Is the short-latency dopamine response too short to signal reward error?" In: *Trends Neurosci* 22.4 (1999), pp. 146–151.

[Red+99b] Peter Redgrave, Tony J Prescott, and Kevin Gurney. "The basal ganglia: a vertebrate solution to the selection problem?" In: *Neuroscience* 89.4 (1999), pp. 1009–1023.

[Red04] A David Redish. "Addiction as a computational process gone awry". In: *Science* 306.5703 (2004), pp. 1944–1947.

[Ren+14] Erwan Renaudo, Benoît Girard, Raja Chatila, and Mehdi Khamassi. "Design of a control architecture for habit learning in robots". In: *Living Machines 2014, Lecture Notes in Artificial Intelligence*. 2014, to appear.

[Res04] Robert A Rescorla. "Spontaneous recovery". In: *Learn Mem* 11.5 (2004), pp. 501–509.

[Res97] Robert A Rescorla. "Spontaneous recovery after Pavlovian conditioning with multiple outcomes". In: *Anim Learn Behav* 25.1 (1997), pp. 99–107.

[Res99] Robert A Rescorla. "Summation and overexpectation with qualitatively different outcomes". In: *Anim Learn Behav* 27.1 (1999), pp. 50–62.

[Rey61] George S Reynolds. "Attention in the pigeon". In: *J Exp Anal Behav* 4.3 (1961), pp. 203–208.

[RF+11] Jose JF Ribas-Fernandes, Alec Solway, Carlos Diuk, Joseph T McGuire, Andrew G Barto, Yael Niv, and Matthew M Botvinick. "A neural signature of hierarchical reinforcement learning". In: *Neuron* 71.2 (2011), pp. 370–379.

[RF09] Terry E Robinson and Shelly B Flagel. "Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences". In: *Biol Psychiatry* 65.10 (2009), pp. 869–873.

[RH75] Robert A Rescorla and C Donald Heth. "Reinstatement of fear to an extinguished conditioned stimulus." In: *J Exp Psychol Anim Behav Process* 1.1 (1975), p. 88.

[Rob+06] Siobhan Robinson, Bethany N Sotak, Matthew J During, and Richard D Palmiter. "Local dopamine production in the dorsal striatum restores goal-directed behavior in dopamine-deficient mice." In: *Behav Neurosci* 120.1 (2006), p. 196.

[Roe+07] Matthew R Roesch, Donna J Calu, and Geoffrey Schoenbaum. "Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards". In: *Nat Neurosci* 10.12 (2007), pp. 1615–1624.

[Roe+12] Matthew R Roesch, Guillem R Esber, Jian Li, Nathaniel D Daw, and Geoffrey Schoenbaum. “Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain”. In: *Eur J Neurosci* 35.7 (2012), pp. 1190–1200.

[Ros+07] Juan M Rosas, Ana García-Gutiérrez, and José E Callejas-Aguilera. “AAB and ABA Renewal as a Function of the Number of Extinction Trials in Conditioned Taste Aversion.” In: *Psicologica: International Journal of Methodology and Experimental Psychology* 28.2 (2007), pp. 129–150.

[RR72] Ross C Rizley and Robert A Rescorla. “Associations in second-order conditioning and sensory preconditioning.” In: *J Comp Physiol Psychol* 81.1 (1972), p. 1.

[RW72] Robert A Rescorla and Allan R Wagner. “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement”. In: *Classical conditioning II: Current research and theory* 2 (1972), pp. 64–99.

[Sam+05] Kazuyuki Samejima, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. “Representation of action-specific reward values in the striatum”. In: *Science* 310.5752 (2005), pp. 1337–1340.

[San+06] Federico Sanabria, Matthew T Sitomer, and Peter R Killeen. “Negative automaintenance omission training is effective”. In: *J Exp Anal Behav* 86.1 (2006), pp. 1–10.

[SB12] Alec Solway and Matthew M Botvinick. “Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates”. In: *Psychol Rev* 119.1 (2012), pp. 120–154.

[SB81] Richard S Sutton and Andrew G Barto. “Toward a modern theory of adaptive networks: expectation and prediction.” In: *Psychol Rev* 88.2 (1981), p. 135.

[SB87] Richard S Sutton and Andrew G Barto. “A temporal-difference model of classical conditioning”. In: *Proceedings of the ninth annual conference of the cognitive science society.* Seattle, WA. 1987, pp. 355–378.

[SB90] Richard S Sutton and Andrew G Barto. “Time-derivative models of pavlovian reinforcement.” In: *Learning and computational neuroscience: Foundations of adaptive networks.* The MIT Press, 1990, pp. 497–537.

[SB98] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* The MIT Press, 1998.

[SC02] John D Salamone and Mercè Correa. “Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine”. In: *Behav Brain Sci* 137.1 (2002), pp. 3–25.

[SC12] John D Salamone and Mercè Correa. “The mysterious motivational functions of mesolimbic dopamine”. In: *Neuron* 76.3 (2012), pp. 470–485.

[Sch+96] Néstor A Schmajuk, Ying-Wan Lam, and J A Gray. “Latent inhibition: A neural network approach.” In: *J Exp Psychol Anim Behav Process* 22.3 (1996), pp. 321–349.

[Sch+97] Wolfram Schultz, Peter Dayan, and P Read Montague. “A neural substrate of prediction and reward”. In: *Science* 275.5306 (1997), pp. 1593–1599.

[Sch07] Wolfram Schultz. “Behavioral dopamine signals”. In: *Trends Neurosci* 30.5 (2007), pp. 203–210.

[Sch10] Wolfram Schultz. “Review Dopamine signals for reward value and risk: basic and recent data”. In: *Behav Brain Funct* 6 (2010), p. 24.

[Sch13] Wolfram Schultz. “Updating dopamine reward signals”. In: *Curr Opin Neurobiol* 23.2 (2013), pp. 229–238.

[Sch98] Wolfram Schultz. “Predictive Reward Signal of Dopamine Neurons”. In: *J Neurophysiol* 80 (1998), pp. 1–27.

[SD12] Dylan A Simon and Nathaniel D Daw. “Dual-system learning models and drugs of abuse”. In: *Computational Neuroscience of Drug Addiction*. Springer, 2012, pp. 145–161.

[SH04] Brian Sallans and Geoffrey E Hinton. “Reinforcement learning with factored states and actions”. In: *The Journal of Machine Learning Research* 5 (2004), pp. 1063–1088.

[She+11] Denis Sheynikhovich, Satoru Otani, and Angelo Arleo. “The role of tonic and phasic dopamine for long-term synaptic plasticity in the prefrontal cortex: a computational model”. In: *J Physiol Paris* 105.1 (2011), pp. 45–52.

[She65] Fred D Sheffield. “Relation between classical conditioning and instrumental learning”. In: *Classical conditioning. New York: Appleton-Century-Crofts* (1965), pp. 302–322.

[Shu+80] Thomas R Shultz, Diane Wells, and Mario Sarda. “Development of the ability to distinguish intended actions from mistakes, reflexes, and passive movements”. In: *Br J Soc Clin Psychol* 19.4 (1980), pp. 301–310.

[Sin+00] Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. “Convergence results for single-step on-policy reinforcement-learning algorithms”. In: *Mach Learn* 38.3 (2000), pp. 287–308.

[Ska+13] Anya Skatova, Patricia A Chan, and Nathaniel D Daw. “Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task”. In: *Front Hum Neurosci* 7 (2013).

[Ske+14] Ivan Skelin, Rhys Hakstol, Jenn VanOyen, Dominic Mudiayi, Leonardo A Molina, Victoria Holec, Nancy S Hong, David R Euston, Robert J McDonald, and Aaron J Gruber. “Lesions of dorsal striatum eliminate lose-switch responding but not mixed-response strategies in rats”. In: *Eur J Neurosci* 39.10 (2014), pp. 1655–1663.

[Ski38] Burrhus F Skinner. *The behavior of organisms: An experimental analysis*. Appleton-Century-Crofts New York, 1938, pp. 82–82.

[SL06] Nestor A Schmajuk and José A Larrauri. “Experimental challenges to theories of classical conditioning: application of an attentional model of storage and retrieval.” In: *J Exp Psychol Anim Behav Process* 32.1 (2006), p. 1.

[SM07] Steven C Stout and Ralph R Miller. “Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis.” In: *Psychol Rev* 114.3 (2007), pp. 759–783.

[Smi+13] Peter Smittenaar, Thomas HB FitzGerald, Vincenzo Romei, Nicholas D Wright, and Raymond J Dolan. “Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans”. In: *Neuron* 80.4 (2013), pp. 914–919.

[SR12] Benjamin T Saunders and Terry E Robinson. “The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses”. In: *Eur J Neurosci* 36.4 (2012), pp. 2521–2532.

[SR13] Benjamin T Saunders and Terry E Robinson. “Individual variation in resisting temptation: implications for addiction”. In: *Neurosci Biobehav Rev* 37.9 (2013), pp. 1955–1975.

[Ste+13] Elizabeth E Steinberg, Ronald Keiflin, Josiah R Boivin, Ilana B Witten, Karl Deisseroth, and Patricia H Janak. “A causal link between prediction errors, dopamine neurons and learning”. In: *Nat Neurosci* 16.7 (2013), pp. 966–973.

[Sut+92] Richard S Sutton, Andrew G Barto, and Ronald J Williams. “Reinforcement learning is direct adaptive optimal control”. In: *Control Systems, IEEE* 12.2 (1992), pp. 19–22.

[Sut+99] Richard S Sutton, Doina Precup, and Satinder Singh. “Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning”. In: *Artificial intelligence* 112.1 (1999), pp. 181–211.

[Sut88] Richard S Sutton. “Learning to predict by the methods of temporal differences”. In: *Mach Learn* 3.1 (1988), pp. 9–44.

[Sut90] Richard S Sutton. “Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming.” In: *ML*. 1990, pp. 216–224.

[SW09] Lars Schwabe and Oliver T Wolf. “Stress prompts habit behavior in humans”. In: *J Neurosci* 29.22 (2009), pp. 7191–7198.

[SW11] Lars Schwabe and Oliver T Wolf. “Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action”. In: *Behav Brain Sci* 219.2 (2011), pp. 321–328.

[Sze10] Csaba Szepesvári. “Algorithms for reinforcement learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 4.1 (2010), pp. 1–103.

[Tak+09] Yuji K Takahashi, Matthew R Roesch, Thomas A Stalnaker, Richard Z Haney, Donna J Calu, Adam R Taylor, Kathryn A Burke, and Geoffrey Schoenbaum. “The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes”. In: *Neuron* 62.2 (2009), pp. 269–280.

[Tal+08] Deborah Talmi, Ben Seymour, Peter Dayan, and Raymond J Dolan. “Human Pavlovian–instrumental transfer”. In: *J Neurosci* 28.2 (2008), pp. 360–368.

[Tho+03] Brian L Thomas, Niccole Larsen, and John JB Ayres. “Role of context similarity in ABA, ABC, and AAB renewal paradigms: Implications for theories of renewal and for treating human phobias”. In: *Learn Motiv* 34.4 (2003), pp. 410–436.

[Tho+10] Catherine A Thorn, Hisham Atallah, Mark Howe, and Ann M Graybiel. “Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning”. In: *Neuron* 66.5 (2010), pp. 781–795.

[Tho11] Edward L Thorndike. *Animal Intelligence.* 1911.

[Tin+09] Amy J Tindell, Kyle S Smith, Kent C Berridge, and J Wayne Aldridge. "Dynamic computation of incentive salience:"wanting" what was never "liked"". In: *J Neurosci* 29.39 (2009), pp. 12220–12228.

[Tod+12] Travis P Todd, Neil E Winterbauer, and Mark E Bouton. "Effects of the amount of acquisition and contextual generalization on the renewal of instrumental behavior after extinction". In: *Learn Behav* 40.2 (2012), pp. 145–157.

[Tol38] Edward C Tolman. "The determiners of behavior at a choice point." In: *Psychol Rev* 45.1 (1938), p. 1.

[Tol39] Edaward Chace Tolman. "Prediction of vicarious trial and error by means of the schematic sowbug." In: *Psychol Rev* 46.4 (1939), p. 318.

[Tom+03] Arthur Tomie, Eugene D Festa, Dennis R Sparta, and Larissa A Pohorecky. "Lever conditioned stimulus–directed autoshaping induced by saccharin–ethanol unconditioned stimulus solution: effects of ethanol concentration and trial spacing". In: *Alcohol* 30.1 (2003), pp. 35–44.

[Tom+12] Arthur Tomie, Michelle Lincks, Steffi D Nadarajah, Larissa A Pohorecky, and Lei Yu. "Pairings of lever and food induce Pavlovian conditioned approach of sign-tracking and goal-tracking in C57BL/6 mice". In: *Behav Brain Res* 226.2 (2012), pp. 571–578.

[Tri+09] Elizabeth Tricomi, Bernard W Balleine, and John P O'Doherty. "A specific role for posterior dorsolateral striatum in human habit learning". In: *Eur J Neurosci* 29.11 (2009), pp. 2225–2232.

[TS12] Nicolas X Tritsch and Bernardo L Sabatini. "Dopaminergic modulation of synaptic transmission in cortex and striatum". In: *Neuron* 76.1 (2012), pp. 33–50.

[Val+07] Vivian V Valentin, Anthony Dickinson, and John P O'Doherty. "Determining the neural substrates of goal-directed learning in the human brain". In: *J Neurosci* 27.15 (2007), pp. 4019–4026.

[VB08] Christopher M Vigorito and Andrew G Barto. "Autonomous hierarchical skill acquisition in factored mdps". In: *Yale Workshop on Adaptive and Learning Systems, New Haven, Connecticut.* Vol. 63. 95. 2008, p. 109.

[VN00] Coryn Vickrey and Allen Neuringer. "Pigeon reaction time, Hick's law, and intelligence". In: *Psychon Bull Rev* 7.2 (2000), pp. 284–291.

[Wal+10] Thomas J Walsh, Sergiu Goschin, and Michael L Littman. "Integrating Sample-Based Planning and Model-Based Reinforcement Learning." In: *AAAI.* 2010.

[WD92] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Mach Learn* 8.3-4 (1992), pp. 279–292.

[Wel80] AT Welford. "Choice reaction time: Basic concepts". In: *Reaction times* (1980), pp. 73–128.

[Wie+13] Heather M Wied, Joshua L Jones, Nisha K Cooch, Benjamin A Berg, and Geoffrey Schoenbaum. "Disruption of model-based behavior and learning by cocaine self-administration in rats". In: *Psychopharmacology* 229.3 (2013), pp. 493–501.

[Wil+12] Brian J Wiltgen, Courtney Sinclair, Chadrick Lane, Frank Barrows, Martín Molina, and Chloe Chabanon-Hicks. “The effect of ratio and interval training on Pavlovian-instrumental transfer in mice”. In: *PLoS One* 7.10 (2012), e48227.

[Wil94a] Ben A Williams. “Blocking despite changes in reinforcer identity”. In: *Anim Learn Behav* 22.4 (1994), pp. 442–457.

[Wil94b] Ben A Williams. “Conditioned reinforcement: Experimental and theoretical issues”. In: *Behav Anal* 17.2 (1994), pp. 261–285.

[Wit+09] Sanne de Wit, Philip R Corlett, Mike R Aitken, Anthony Dickinson, and Paul C Fletcher. “Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans”. In: *J Neurosci* 29.36 (2009), pp. 11330–11338.

[WO12] Marco Wiering and Martijn van Otterlo. *Reinforcement Learning*. Springer, 2012.

[Woo+74] William T Woodard, John C Ballinger, and ME Bitterman. “Autoshaping: further study of ”negative automaintenance””. In: *J Exp Anal Behav* 22.1 (1974), pp. 47–51.

[Wun+12] Klaus Wunderlich, Peter Smittenaar, and Raymond J Dolan. “Dopamine enhances model-based over model-free choice behavior”. In: *Neuron* 75.3 (2012), pp. 418–424.

[WW69] David R Williams and Harriet Williams. “Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement.” In: *J Exp Anal Behav* (1969).

[Yam+13] Hiroshi Yamada, Hitoshi Inokawa, Naoyuki Matsumoto, Yasumasa Ueda, Kazuki Enomoto, and Minoru Kimura. “Coding of the long-term value of multiple future rewards in the primate striatum”. In: *J Neurophysiol* 109.4 (2013), pp. 1140–1151.

[Yin+04] Henry H Yin, Barbara J Knowlton, and Bernard W Balleine. “Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning”. In: *Eur J Neurosci* 19.1 (2004), pp. 181–189.

[Yin+05] Henry H Yin, Sean B Ostlund, Barbara J Knowlton, and Bernard W Balleine. “The role of the dorsomedial striatum in instrumental conditioning”. In: *Eur J Neurosci* 22.2 (2005), pp. 513–523.

[Yin+06] Henry H Yin, Barbara J Knowlton, and Bernard W Balleine. “Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning”. In: *Behav Brain Sci* 166.2 (2006), pp. 189–196.

[Yin+08] Henry H Yin, Sean B Ostlund, and Bernard W Balleine. “Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks.” In: *Eur J Neurosci* 28.8 (2008), pp. 1437–1448.

[YK06] Henry H Yin and Barbara J Knowlton. “The role of the basal ganglia in habit formation”. In: *Nat Rev Neurosci* 7.6 (2006), pp. 464–476.

[Zha+09] Jun Zhang, Kent C Berridge, Amy J Tindell, Kyle S Smith, and J Wayne Aldridge. “A neural computational model of incentive salience”. In: *PLoS Comput Biol* 5.7 (2009), e1000437.

[Bai93] Lemon C Baird III. *Advantage Updating.* Tech. rep. DTIC Document, 1993.