Interactions entre circuits préfrontaux et neurones dopaminergiques dans la prise de décision chez la souris

Elise Bousseyrol

▶ **To cite this version:**

Elise Bousseyrol. Interactions entre circuits préfrontaux et neurones dopaminergiques dans la prise de décision chez la souris. Neurosciences [q-bio.NC]. Sorbonne Université, 2021. Français. ⟨NNT : 2021SORUS208⟩. ⟨tel-03481993⟩

HAL Id: tel-03481993

https://theses.hal.science/tel-03481993v1

Submitted on 15 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Thèse de Doctorat de Sorbonne Université

Spécialité Neurosciences

École Doctorale Cerveau – Cognition – Comportement

Pour obtenir le grade de

Docteur de Sorbonne Université

Présentée par

Élise Bousseyrol

Sujet de la thèse :

Interactions entre circuits préfrontaux et neurones dopaminergiques dans la prise de décision chez la souris

Soutenue publiquement le 19 Novembre 2021

Devant un jury composé de :

Dr. Stéphanie TROUCHE	Rapporteure
Dr. David ROBBE	Rapporteur
Dr. Guillaume DUGUÉ	Examinateur
Dr. Gabrielle GIRARDEAU	Examinatrice
Dr. Emmanuel PROCYK	Président du jury
Dr. Philippe FAURE	Directeur de thèse
Dr. Jérémie NAUDÉ	Invité

Sorbonne Université
Bureau d'accueil des doctorants
Institut de Formation Doctorale
scolarite.doctorat@sorbonne-universite.fr

Les Cordeliers
15 rue de l'Ecole de Médecine
75006 – Paris
Escalier G, 2ème étage

Remerciements

Table des matières

Table des matières 3
Tables des figures 7
Résumé 9
Abstract 11
Liste des publications 13
Liste des abréviations 15
Introduction générale 19
Chapitre 1 : Récompense et comportements de prise de décision 23
I) Les effets comportementaux d'une récompense 23
1) Récompense et hédonisme 23
2) Récompense et conditionnements 26
3) Récompense et motivation vers un but 33
II) Différents systèmes d'attribution de valeur 36
1) Le système pavlovien 36
2) Le système orienté vers un but 38
3) Le système habituel 40
4) Interactions entre systèmes d'attribution de valeur 41
III) Prendre une décision 42
1) Choix économique 43
2) Valeur et utilité 44
3) Politique 48
IV) Balance exploitation / exploration 49
1) L'exploration modulée par des paramètres externes 49
2) L'exploration dans la sélection de l'action 50
V) Conclusions 53

Chapitre 2 : Le circuit méso-cortico-limbique 55

I) Le système dopaminergique 55

1) Dopamine et neurotransmission 55

2) Anatomie des voies dopaminergiques de l'Aire Tegmentale Ventrale 57

3) Activité spontanée des neurones dopaminergiques et relargage 64

II) Dopamine, récompense et prise de décision 69

1) Le système dopaminergique et la stimulation intracrânienne 69

2) L'erreur de prédiction de récompense 70

3) L'erreur de prédiction de récompense comme signal d'apprentissage par renforcement 72

4) L'erreur de prédiction de récompense comme signal quantitatif de valeur 74

5) Incertitude liée à la récompense 76

6) L'erreur de prédiction de récompense comme monnaie commune 77

7) Neuromodulation dopaminergique et prise de décision 79

Chapitre 3 : Les cortex préfrontaux 81

I) Organisation des cortex préfrontaux 81

1) Architecture des cortex préfrontaux 81

2) Activité oscillatoire des cortex 84

II) Rôles du Cortex Orbitofrontal et Préfrontal dans la prise de décision 90

3) Le Cortex Orbitofrontal 90

4) Le Cortex Préfrontal 93

III) Interactions entre l'Aire Tegmentale Ventrale et le Cortex Orbitofrontal et Préfrontal 97

Chapitre 4 : Paradigme comportemental 99

I) La tâche 3 points 99

1) Principe de la tâche 99

2) Variations de la tâche 100
II) Comportement des souris dans la tâche 102
1) L'essai comme comportement de base de la tâche 102
2) Comportement locomoteur dans la tâche 103
3) Incertitude et balance Exploration / Exploitation 105
III) Modélisation des choix dans la tâche 107
1) Le modèle 107
2) L'estimation des paramètres du modèle 109
IV) Enregistrements électrophysiologiques chez l'animal éveillé dans la tâche 112
1) Activité multi-unitaire des neurones dopaminergiques 112
2) Activité d'ensemble des cortex préfrontaux 114
Chapitre 5 : Première publication 117
Chapitre 6 : Deuxième publication 163
Chapitre 7 : Discussion générale 181
I) Activités prédictives ou causales 182
1) Activité dopaminergique et erreur de prédiction de récompense 183
2) Activité dans l'OFC et la représentation d'état 185
3) Le cortex préfrontal et le changement de contingences 186
II) Séquence d'activités et contextes 187
1) Des activités organisées en séquence 188
2) Dépendance au contexte 189
III) L'exploration dans la prise de décision 193
1) La recherche d'incertitude 193
2) La composante motivationnelle de l'exploration 195
Annexe 1 – Dongelmans et al. 2020 199
Bibliographie 251

Tables des figures

FIGURE 1 : Conditionnement animal 28
FIGURE 2 : Deux systèmes de comportements volontaires 39
FIGURE 3 : Fonctions d'utilité et attitude face au risque 47
FIGURE 4 : Synthèse et signalisation dopaminergique 56
FIGURE 5 : Neuroanatomie de la VTA et du circuit méso-cortico-limbique 58
FIGURE 6 : Activité spontanée des neurones dopaminergiques de la VTA 66
FIGURE 7 : Les neurones dopaminergiques encodent une RPE 73
FIGURE 8 : Corrélats de valeur et d'incertitude dans l'activité dopaminergique 75
FIGURE 9 : Anatomie et organisation des cortex préfrontaux 83
FIGURE 10 : Dipôles et oscillations 87
FIGURE 11 : Plusieurs encodages dans l'OFC 92
FIGURE 12 : Encodages dans le PFC et changement de stratégie 95
FIGURE 13 : Les différentes variantes de la tâche 3 points 101
FIGURE 14 : Comportement locomoteur dans la tâche 3 points 104
FIGURE 15 : Modélisation des décisions en cadre probabiliste 110
FIGURE 16 : Analyse spectrale par transformée en ondelettes de Morlet 115
FIGURE 17 : La séquence d'activités mésocorticales dépend du contexte 191
FIGURE 18 : Différents substrats du contrôle des décisions dans la tâche 197

Résumé

Les comportements de prise de décision dirigés vers une récompense reposent majoritairement sur la voie méso-cortico-limbique composée de l'Aire Tegmentale Ventrale (VTA), du Noyau Accumbens et des cortex préfrontaux (OFC et PFC), également appelée « circuit de la récompense ». Ce circuit a été impliqué dans *i)* la représentation des contingences de l'environnement et la prédiction des conséquences d'un stimulus ou d'une action, *ii)* l'encodage de valeur, *iii)* l'initiation et la planification de comportements dirigés vers un but, et *iv)* la mise en place de stratégies cognitives adaptatives et flexibles. Ce système optimiserait en particulier les choix futurs vers la répétition d'actions associées aux options de plus haute valeur, ce que l'on définit comme l'exploitation. Cependant, les individus s'engagent spontanément et fréquemment dans des actions qui ne mènent pas forcément à des récompenses ou qui semblent sous-optimales, en particulier lorsque l'environnement est incertain, complexe ou changeant. On parle alors d'exploration. Les comportements de choix peuvent découler de stimuli externes ou de représentations internes à l'animal. Ces représentations peuvent être celles de la récompense potentielle ou de l'information à gagner en lien avec une option. Par ailleurs, l'exploration peut être dirigée vers une option informative ou dériver d'une stratégie de décision plus ou moins aléatoire. Il est difficile d'étudier expérimentalement les choix auto-générés pour en comprendre les bases neuronales, en particulier lorsqu'ils sont variables et exploratoires.

Au cours de ma thèse, j'ai travaillé sur une tâche dans laquelle les souris réalisent des séquences de choix binaires successifs en se basant sur une estimation de contingences afin d'obtenir des récompenses. Dans cette tâche, les choix sont auto-générés puisque les animaux déterminent eux-mêmes l'initiation, la direction, la vigueur et la structure temporelle de leurs actions. En couplant des enregistrements électrophysiologiques et des expériences de perturbation, nous avons identifié une séquence d'activités oscillatoires et de décharge des neurones dopaminergiques de la VTA, de l'OFC et du PFC qui co-encode et co-détermine à la fois le but, l'initiation et le rythme d'exécution de l'action ainsi que la stratégie de choix.

Par ailleurs, l'utilisation d'une règle récompensant les souris lorsqu'elles génèrent des séquences variables, nous a permis de découvrir que les souris sont capables d'accroitre stratégiquement la variabilité dans leurs séquences de choix en renforçant une stratégie aléatoire, sans se baser sur la mémoire des choix précédents.

Abstract

Reward-guided decision-making behaviour relies mostly on the meso-cortico-limbic pathway, composed of the Ventral Tegmental Area (VTA), the Nucleus Accumbens (NAc) and prefrontal cortices (OFC and PFC), also called "reward circuit". This circuit has been implicated in i) the encoding of value, ii) the representation of contingencies and the prediction of a stimulus' or an action's outcome, iii) the initiation and planification or goal-directed actions, iv) and the implementation of adaptive and flexible cognitive strategies to optimise choice behaviours. This system bias future choices towards the repetition of highest value options, a behavior defined as exploitation. However, individuals engage spontaneously and frequently, in actions and behaviours that do not necessarily lead to reward or that seem sub-optimal, especially in uncertain, complex or changing environments. This is called exploration. Choice behaviours can result from external stimuli or from dynamic and internal representations of the animal. It is difficult to distinguish experimentally between the different motivational factors that lead to variability and exploration in choices, especially when actions are self-paced.

During my thesis, I used a behavioural task in which mice self-determine the initiation, direction, vigor and pace of actions based on contingencies estimation in order to gather rewards in series of successive binary choices. Such goal-directed choices are not triggered by external cues. By using a probabilistic rule to deliver reward coupled with electrophysiological recordings and perturbations, we identified a sequence of oscillations and firing in VTA DAergic neurons, OFC and PFC that co-encodes and co-determines both the goal, initiation and pace of actions as well as choices. Furthermore, by using a reward delivery rule where mice are rewarded for non-repetitive choice sequences, we found that mice progressively increased their choice variability and reinforced a pseudo-random selection strategy, without relying on the memory of previous choices.

Liste des publications

Ce travail de thèse a été effectué sous la supervision de Dr. Philippe Faure, à l'Institut de Biologie Paris Seine, sur le campus Pierre et Marie Curie de Sorbonne Université, au sein de l'unité Neurosciences Paris Seine dirigée par le Dr. Hervé Chneiweiss. Mon travail dans l'équipe Neurophysiologie et Comportement dirigée par Dr. Philippe Faure et Dr. Alexandre Mourot, a été co-encadré par Dr. Jérémie Naudé, et a donné lieu à deux publications, une en tant que première auteure en préparation et une en tant que seconde auteure publiée en 2020 :

- "Self-generation of goal-directed choices in a distributed dopaminergic and prefrontal circuit". **Bousseyrol E**[1,2#], Didienne S[1,2#], Takillah S[1,2], Prévost-Solié C[1,2], Come M[1,2], Ahmed Yahia T[1,2], Tricoire L[2], Dalkara D[3], Mourot A[1,2], Naudé J[1,2†*] & Faure Ph[1,2†*]. In prep.
- "Mice adaptively generate choice variability in a deterministic task". Belkaid M, **Bousseyrol E**, Durand-de Cuttoli R, Dongelmans M, Duranté E K, Ahmed Yahia T, Didienne S, Hannesse B, Come M, Mourot A, Naudé J, Sigaud O* & Faure P*. Communications Biology. 3, 34 (2020).

Les collaborations au sein de l'équipe ont également permis une publication en tant que co-auteure :

- Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in a foraging task. Dongelmans M*, Durand-de Cuttoli R*, Nguyen C*, Come M*, Durante E K, Lemoine D, Brito R, Ahmed Yahia T, Mondoloni S, Didienne S, **Bousseyrol E**, Hannesse B, Reynolds L, Torquet N, Dalkara D, Marti F, Mourot A, Naudé J, Faure P. Nature Communications (2021, in press).

Liste des abréviations

AC Adénylate Cyclase

ACA Aire Cingulaire Antérieure

ACC Cortex Antérieur Cingulaire

ACd Cortex Antérieur dorsal

ACh Acétylcholine

AHP Post Hyperpolarisation (*AfterHyperPolarisation*)

Al Aire Insulaire agranulaire

Amg Amygdale

AMPc Adénosine Monophosphate cyclique

ATP Adénosine Triphosphate

BLA Amygdale Basolatérale

CHR2 ChannelRhodopsine - 2

CLi noyau Linéaire Caudal

CP Pédoncule Cérébelleux supérieur

CPP Conditionnement de Préférence de Place

CS Stimulus Conditionné (*Conditionned Stimulus*)

COMT Catécholamine-O-Méthyl-Transférase

DA Dopamine

DAT Transporteur de la Dopamine

DDC Dopamine Décarboxylase

DP noyau Pédonculaire Dorsal

EFP Potentiel de Champ Extracellulaire (*Extracellular Field Potential*)

EPSP Potentiels Post-Synaptiques Excitateurs

Fr Fornix

GABA Acide γ-Aminobutyrique

GAD GABA Décarboxylase

GAT Transporteur du GABA

GIRK canal Potassique à Rectification Entrante activé par les protéines G (*G protein-coupled Inwardly Rectifying Potassium channel*)

GLU Glutamate

GR Récepteur nucléaire aux Glucocorticoïdes

HFHB Haute Fréquence de décharge et Haute fréquence de Bouffées (*High Frequency and High Burst firing*)

HFLB Haute Fréquence de décharge et Basse fréquence de Bouffées (*High Frequency and Low Burst firing*)

HIpp Hippocampe

ICSS Autostimulation Intracrânienne (*Intracranial Self Stimulation*)

IF noyau Interfasciculaire

IL Infralimbique

ILA Aire Infralimbique

IPN Noyau Interpédonculaire

IPSP Potentiels Post-Synaptiques Inhibiteurs

Kv canaux Potassiques Voltage-dépendants

LDTg noyau Tegmental Latéro-Dorsal

LFHB Basse Fréquence de décharge et Haute fréquence de Bouffées (*Low Frequency and High Burst firing*)

LFLB Basse Fréquence de décharge et Basse fréquence de Bouffées (*Low Frequency and Low Burst firing*)

LFP Potentiel de Champ Local (*Local Field Potential*)

LH Hypothalamus Latéral

LHb Habénula Latérale

LO Orbitofrontal Latéral

MFB Faisceau Médian du Télencéphable (*Medial Forebrain Bundle*)

ML Lemnisque Médian

MP Pédoncule Mamillaire

MOs aire Motrice secondaire

mPFC Cortex Préfrontal médian

MSN Neurones Epieux Moyens (*Medium Spiny Neurons*)

MUA Activité Multi-Unitaire

NAc Noyau Accumbens

nAChR Récepteur nicotinique à l'Acétylcholine

OFC Cortex Orbitofrontal

ORB aire Orbitofrontale

PA Potentiel d'Action

PBP noyau Parabrachial Pigmenté

PFC Cortex Préfrontal

PKA Protéine Kinase A

PL aire Prélimbique

PN Noyau latéral Parabrachial

PPTg Noyau Tegmental Pédonculo-Pontin

PV Parvalbumine

R Récompense

RLi noyau Linéaire Rostral

RMTg Tegmentum RostroMédial

RPE Erreur de Prédiction de Récompense

RRF Champ Rétro-Rubral (*Retro Rubral Field*)

SNc Substance Noire *pars compacta*

SNR Substance Noire *pars Reticulata*

TANs Neurones Toniques Actifs (*Tonically Active Neurons*)

TD Trace temporelle (*Temporal Difference*)

TH Tyrosine Hydroxylase

US Stimulus Inconditionné (*Unconditionned Stimulus*)

VGCC Canaux Calciques Voltage-dépendants (*Voltage-Gated Calcium Channels*)

VGLUT Transporteur Vésiculaire du Glutamate

vlPFC Cortex Préfrontal ventrolatéral

VMAT Transporteur Vésiculaire des Monoamines

vmPFC Cortex Préfrontal ventro-médian

VO Orbitofrontal Ventral

VTA Aire Tegmentale Ventrale

WT Type Sauvage (*Wild Type*)

Introduction générale

Que ce soit choisir quelle paire de chaussettes mettre le matin, ce que l'on mange au déjeuner, dans quel logement vivre, ou bien comment organiser nos journées de travail, nos grandes ou petites décisions construisent nos vies. Quand un individu prend une décision, il dispose de plus au moins d'informations à la fois sur son état interne, le contexte dans lequel il se trouve et son but, afin de sélectionner un plan d'action approprié. Le cerveau semble construire, par l'expérience, une représentation mentale de l'environnement et des associations entre les actions et leur résultat (*outcome*), qui sert à calculer le bénéfice attendu de chaque option, puis à les comparer et choisir en appliquant une politique (ou stratégie). Le bénéfice peut être une récompense (de nature ou dimension variées), mais aussi un gain d'information lié à l'exploration d'options nouvelles ou à l'issue incertaine. En outre, de nombreux choix dépendent du contexte environnemental, de l'historique récent et de l'homéostasie globale de l'individu. L'objectif général de cette thèse est d'étudier les bases neurales qui sous-tendent les décisions et les stratégies de choix chez la souris.

Le cerveau intègre nombre des éléments influençant les valeurs des options ou les stratégies de choix, notamment via le circuit méso-cortico-limbique ou « circuit de la récompense ». Ce circuit comporte des structures mésencéphaliques, en particulier l'Aire Tegmentale Ventrale (VTA), composée notamment de neurones dopaminergiques (DAergiques), du striatum qui fait partie du système limbique et en particulier sa partie ventrale appelée Noyau Accumbens (NAc), ainsi que de structures corticales notamment les cortex préfrontal (PFC) et orbitofrontal (OFC). Ce circuit a été impliqué *i)* dans l'encodage de la valeur de stimuli récompensants, *ii)* dans la représentation des contingences de l'environnement et la prédiction de l'*outcome* d'un stimulus ou d'une action, *iii)* dans l'initiation et la planification de comportements motivés dirigés vers un but, *iv)* ainsi que dans la mise en place de stratégies cognitives adaptatives et flexibles pour optimiser les comportements de prise de décision dans des environnements incertains, complexes ou changeants. L'activité des structures dans ces circuits est souvent décrite indépendamment les unes des autres. Nous avons tiré profit de la structure particulière d'une tâche nous avons mise en place pour explorer le rôle respectif de ces différentes structures et leurs interactions.

L'utilisation d'un stimulus permet très souvent de spécifier le moment d'agir ou encore l'option à choisir. Cependant, les animaux (humains inclus) s'engagent spontanément et très fréquemment, sans indice déclencheur, dans des comportements motivés ne menant pourtant

pas forcément à des récompenses. Dans cette thèse, nous développons une tâche dans laquelle les souris déterminent elles-mêmes l'initiation, la direction, la vigueur et le rythme des actions qu'elles choisissent, on parle de comportements auto-générés. Enfin, confronté à un choix binaire entre deux options de valeurs différentes, les individus choisissent une portion significative du temps des options qui paraissent sous-optimales, en particulier si elles sont nouvelles, incertaines ou informatives, ou si l'environnement est complexe ou changeant. C'est ce qu'on appelle l'exploration. Les mécanismes neurophysiologiques permettant d'ignorer ses préférences et l'expérience passée pour explorer, soit de manière dirigée vers les options moins connues ou plus incertaines de l'environnement, soit même de façon aléatoire, sont à ce jour mal décrits. L'encodage de la valeur des options tournées vers l'exploration, et cela d'autant plus dans le cadre de comportements auto-générés, est une notion encore mal comprise. De même, il est difficile de distinguer les différents facteurs motivationnels poussant à générer de la variabilité et de l'exploration dans les choix.

Au cours de ce travail de thèse, je me suis donc intéressée, chez la souris, aux interactions dynamiques entre les structures du circuit méso-cortico-limbique dans des séquences de choix auto-générés, en particulier quand les récompenses sont incertaines (cf. Article 1). J'ai également questionné la capacité des souris à moduler le degré la variabilité de leurs décisions, en renforçant une stratégie d'exploration basée sur le hasard, dans une tâche complexe bien que déterministe (cf. Article 2). L'objectif général a été de comprendre l'émergence d'exploration et de variabilité dans les choix ainsi que les bases neurales qui sous-tendent ces comportements.

Dans ce manuscrit de thèse, je commence donc par définir un certain nombre de concepts de prise de décision concernant la notion de récompense, en particulier l'attribution de valeur et la balance entre exploitation et exploration (cf. Chapitre 1). Je présente ensuite le système méso-cortico-limbique en me concentrant sur la DA et ses fonctions dans la signalisation de récompense et dans la prise de décision (cf. Chapitre 2). La partie suivante s'intéresse aux cortex préfrontaux (l'OFC et le PFC en particulier), à leur organisation et à leur activité. J'y expose leur rôle dans la prise de décision, notamment en interaction avec la VTA (cf. Chapitre 3). Enfin, j'introduis le paradigme comportemental (une tâche à 3 points basée sur un conditionnement opérant par des récompenses distribuées selon différentes règles) et expérimental (enregistrements électrophysiologiques des neurones DAergiques de la VTA et

du potentiel de champ extracellulaire des cortex préfrontaux) utilisé au cours de ma thèse (cf. Chapitre 4).

Ces quatre chapitres introductifs sont suivis de deux publications principales (cf. Chapitres 5 et 6) ainsi que d'une discussion (cf. Chapitre 7). Ces deux publications utilisent la « tâche 3 points » comme test expérimental des prises de décision. Dans cette tâche, les souris effectuent un nombre suffisant grand et régulier de choix pour étudier de manière robuste : *i)* les variations fines de leur comportement, *ii)* leurs choix en contexte déterministe, probabiliste ou complexe, avec l'utilisation de modèles computationnels pour interpréter les stratégies des animaux, et *iii)* les dynamiques d'activités mésocorticales, par des enregistrements et des manipulations causales.

Chapitre 1 : Récompense et comportements de prise de décision

Qu'il s'agisse de choisir quoi manger à déjeuner, dans quel logement vivre, où passer ses vacances, ou bien comment organiser ses journées, la récompense associée aux conséquences de nos choix est un déterminant majeur qui oriente nos décisions. Quelle que soit sa forme – stimulus, événement, situation, objet, etc. – une récompense est capable de susciter l'envie d'en obtenir encore d'autres. Une récompense revêt ainsi trois principaux aspects : i) elle est associée à une sensation de plaisir ou de satisfaction, physique ou morale, ii) elle peut induire un apprentissage, et iii) elle suscite de la motivation en créant des objectifs (Berridge et al., 2009; Berridge and Robinson, 2003; Schultz, 2015, 2006).

I) Les effets comportementaux d'une récompense

1) Récompense et hédonisme

La première propriété venant à l'esprit lorsqu'on décrit une récompense est son aspect hédonique, c'est-à-dire le plaisir immédiat ou l'émotion positive qu'elle provoque. Différentes formes de plaisir peuvent être induites par des récompenses, comme la sensation de plaisir de manger ou boire, le plaisir sexuel, le plaisir de sociabiliser, ou de consommation d'une drogue. La question du lien entre l'effet hédonique d'une récompense et les processus de survie se pose. Le plaisir est-il crucial, ou bien est-ce un épiphénomène lors de l'obtention d'une récompense ? Une vision maintenant consensuelle est que la sensation de plaisir serait un aspect nécessaire à la sélection de comportements adaptatifs. D'un point de vue évolutif, le cerveau est doté de mécanismes permettant de trouver les récompenses améliorant la fitness (valeur sélective) d'un individu et de sa population (Keramati and Gutkin, 2014; LeDoux, 2012). Ce serait donc la recherche de situations plaisantes qui orienterait les décisions d'un individu et permettrait de générer des comportements adaptés (Pessiglione, 2014, 2009). De cette vision découle de nombreux questionnements sur l'existence d'états émotionnels chez les mammifères, de leurs rôles dans les comportements adaptatifs cruciaux pour la survie et l'apprentissage, ainsi que sur

les circuits cérébraux communs à ces processus (Banich and Floresco, 2019; LeDoux, 2012; Panksepp, 2011).

L'homéostasie est un élément de réponse éclairant le rôle adaptatif de l'hédonisme. Ce concept, formulé par Walter Canon en 1925, désigne l'idée qu'il existe des mécanismes opposés permettant de maintenir une variable physiologique (température corporelle, glycémie, etc.) stable dans une certaine gamme qui est viable pour l'organisme et d'assurer ainsi sa survie (Berridge, 2004; Keramati and Gutkin, 2014). Par exemple, la régulation homéostatique de la glycémie repose sur l'action antagoniste de deux hormones, l'insuline et le glucagon, entrant en jeu lorsque la glycémie dépasse le seuil d'hyper ou d'hypoglycémie. D'une manière générale, préserver son homéostasie requiert la mise en œuvre de comportements spécifiques qui garantissent le maintien d'un certain état interne. Une première vision est de considérer que ces comportements seraient déclenchés par des besoins, lesquels moduleraient l'aspect hédonique des récompenses. Par exemple, la faim et la soif déclencheraient des comportements fondamentaux de recherche de récompenses alimentaires, tels que se nourrir ou boire, en augmentant la valeur hédonique de la nourriture au moment opportun. Néanmoins, de nombreux comportements tournés vers des récompenses sont enclenchés avant qu'un besoin physiologique ne survienne. La gourmandise ou la surconsommation de nourriture en sont les parfaits exemples. Nous avons tous vécu la fameuse situation où nous mangeons des cacahuètes ou du chocolat sans trop pouvoir s'arrêter et sans avoir particulièrement faim. Par ailleurs, bien que la saccharine soit un édulcorant artificiel n'ayant aucune valeur nutritionnelle, les rats la considèrent comme une récompense puisqu'ils la préfèrent à l'eau et sont prêts à travailler et fournir un effort pour en obtenir (Berridge, 2004; Keramati and Gutkin, 2014; Schultz, 2006). Plutôt que de réguler notre homéostasie en réponse à un besoin physiologique, c'est donc la recherche du plaisir en soi qui maintiendrait activement l'homéostasie en nous poussant vers des récompenses (on parle de *drive* homéostatique).

Le plaisir déclenché par une récompense, le « plaisir en soi », est difficilement mesurable de manière dissociée d'un comportement. En effet, on peut se demander à quel moment l'hédonisme entre-t-il en jeu ? Est-ce la vision d'une part de gâteau, son goût en bouche, ou l'augmentation de sucre dans le sang suivant sa consommation qui provoque du plaisir ? S'il est possible de demander à des sujets humains d'évaluer le plaisir qu'ils ressentent en obtenant une récompense, chez l'animal, cela nécessite souvent de mesurer la motivation à effectuer une

tâche menant à cette récompense : il n'est alors pas évident de dissocier l'aspect hédonique d'une récompense du comportement qui permet de l'obtenir.

Dans les années 1960, Young s'est intéressé aux procédés hédoniques en étudiant des comportements de préférence alimentaire chez des rats (Stellar, 1982; Young, 1966). Il mesure l'intensité hédonique d'une récompense chez des rats en comparant leurs préférences pour des boissons contenant différentes concentrations de sucrose et en mesurant la fréquence de consommation. Pour cela, il entraîne les animaux à choisir parmi 2 ou 8 bouteilles lors de tests de durée brève (quelques minutes). Il montre notamment que l'introduction de récompenses à forte valeur positive (c'est-à-dire à haute concentration de sucrose) induisent des changements immédiats de comportements alimentaires : les rats augmentent drastiquement la fréquence de léchage (*licking*) par minute. Ainsi, la fréquence du comportement semble indiquer le plaisir ressenti par l'animal. Mais il souligne néanmoins que le potentiel hédonique d'une récompense ne peut se réduire au comportement de consommation de celle-ci. Il observe en effet que le comportement de léchage des rats atteint une fréquence maximale et stable pour des concentrations de sucrose nettement inférieures aux concentrations préférées par les rats. Ces derniers préfèreront consommer des boissons plus concentrées en sucrose dès que cela leur sera proposé, mais à partir d'un certain seuil, le taux de *licking* n'augmentera pas davantage et ne sera plus corrélé à leurs préférences. Les études de Young soulignent donc que les mesures de performance telle que la fréquence d'un comportement ne sont pas de bons indicateurs d'intensité hédoniques puisqu'elles plafonnent pour de fortes valeurs.

A la fin du 20ème siècle, Kent C. Berridge, développe une théorie différenciant le « *liking* » du « *wanting* » (Berridge, 2004, 1999; Berridge and Robinson, 1998). Selon lui, un contexte associé à une récompense devient apprécié sur le plan hédonique, c'est la notion de *liking*. Ce plaisir ressenti devient ensuite un désir et est activement recherché sur le plan comportemental, c'est la notion de *wanting*. Le plaisir (*liking*) et la recherche du plaisir (*wanting*) sont donc deux concepts distincts qui peuvent être exprimés par un individu indépendamment l'un de l'autre. Cette dissociation est notamment intéressante dans l'étude de l'addiction aux substances d'abus, une fois cette pathologie installée on retrouve des comportements de *wanting* sans la dimension du *liking* initialement induite par la consommation de la drogue (Berridge, 1999). Les méthodes de mesures de l'hédonisme d'une récompense ont très souvent été des quantifications comportementales du *wanting* (mesures de préférences ou de performances) à partir desquelles la composante du *liking* était inférée. Ces méthodes d'évaluation demeurent encore majoritaires en raison de leur facilité de mise en place, mais présentent néanmoins des

limites, comme expliqué précédemment. Berridge propose de se baser sur les réactions affectives des individus (Berridge, 1999; Berridge and Robinson, 1998) pour mesurer l'hédonisme d'une récompense telles que des expressions faciales. Cela peut s'appliquer à la fois chez les jeunes enfants ne parlant pas encore ainsi que chez les animaux, et présente l'avantage de ne pas être lié à une auto-évaluation subjective du sujet. Il faut cependant reconnaître que ce type de mesure est plus évident chez les primates, qui ont des réactions émotionnelles plus proches des nôtres, que chez d'autres mammifères comme les rongeurs. Plusieurs grilles d'évaluations d'expressions faciales et corporelles ont été néanmoins développées afin d'améliorer et compléter les méthodes de quantification de la valeur hédonique des récompenses (Finlayson et al., 2016; Parr et al., 2005).

2) Récompense et conditionnements

Une récompense, parce qu'elle présente un attrait, est capable de déclencher un changement comportemental qui perdure dans le temps. C'est ainsi que le champ de la psychologie définit l'apprentissage (Pessiglione, 2014; Schultz, 2006), et défend l'étude des conséquences visibles de la récompense.

C'est au début du 20ème siècle que naît le courant « behavioriste », courant de psychologie expérimentale souhaitant comprendre les comportements et particulièrement l'apprentissage. Le behaviorisme est fondé en 1913 par John Watson, psychologue américain, qui défend une approche objective et quantifiable de la psychologie, par opposition aux courants mentalistes qui s'intéressaient au fonctionnement de la conscience humaine par le biais de l'introspection. L'étude des associations entre un stimulus et un comportement est au cœur du courant behavioriste, qui ne questionne pas les processus mentaux survenant entre le stimulus et la réponse et exclue toute représentation mentale et abstraite comme facteur causal d'un comportement.

Le courant behavioriste découle des travaux d'Ivan Pavlov[1], physiologiste et médecin russe, qui, dans les années 1890, décrit le premier les étapes du conditionnement simple. Pavlov étudiait alors les fonctions gastriques chez le chien ainsi que la salivation, et remarquait que le réflexe de salivation semblait précéder le moment où il donnait la nourriture au chien. Ce constat l'a mené à mettre au point une expérience de conditionnement classique, dit plus tard de conditionnement pavlovien.

Dans cette expérience (Figure 1A), Pavlov donne de la nourriture à un chien et mesure la salivation induite. Il fait alors retentir une cloche qui précède le moment de la distribution de nourriture au chien. Après répétition, le chien a fait l'association entre le son de cloche et la récompense, et se met à saliver au moment où il entend la cloche. Dans ce paradigme, le réflexe de salivation constitue une réponse inconditionnée, la nourriture est un stimulus inconditionné (US, *unconditionned stimulus* en anglais), et le son de cloche un stimulus conditionné (CS, *conditionned stimulus*). Après conditionnement, le chien salive à chaque fois qu'il entend le son de la cloche, et ce même si la nourriture n'est pas délivrée par la suite : on parle alors de réponse conditionnée. Ivan Pavlov désignera cela sous le nom de réflexe pavlovien, avant que le terme de conditionnement soit plus tard adopté par ses successeurs du mouvement behavioriste. Le conditionnement pavlovien (ou simple ou classique) est donc le fruit d'un apprentissage associatif passif entre un stimulus particulier de l'environnement, qui est au départ neutre, et une réaction innée automatique et non volontaire. Pavlov définit alors une récompense comme un objet produisant un changement comportemental par apprentissage (Schultz, 2015, 2006).

[1] Ivan Pavlov (1849-1936) est un médecin et physiologiste russe dont les idées sont devenues des concepts clés du courant behavioriste. Alors qu'il réalisait des expériences sur la fonction gastrique chez le chien, il découvre et théorise les premières lois fondamentales de l'acquisition de réflexes conditionnés. Suite au caractère novateur de ses travaux, il reçoit le prix Nobel de physiologie et médecine en 1904, et le conditionnement classique est aussi dit pavlovien.

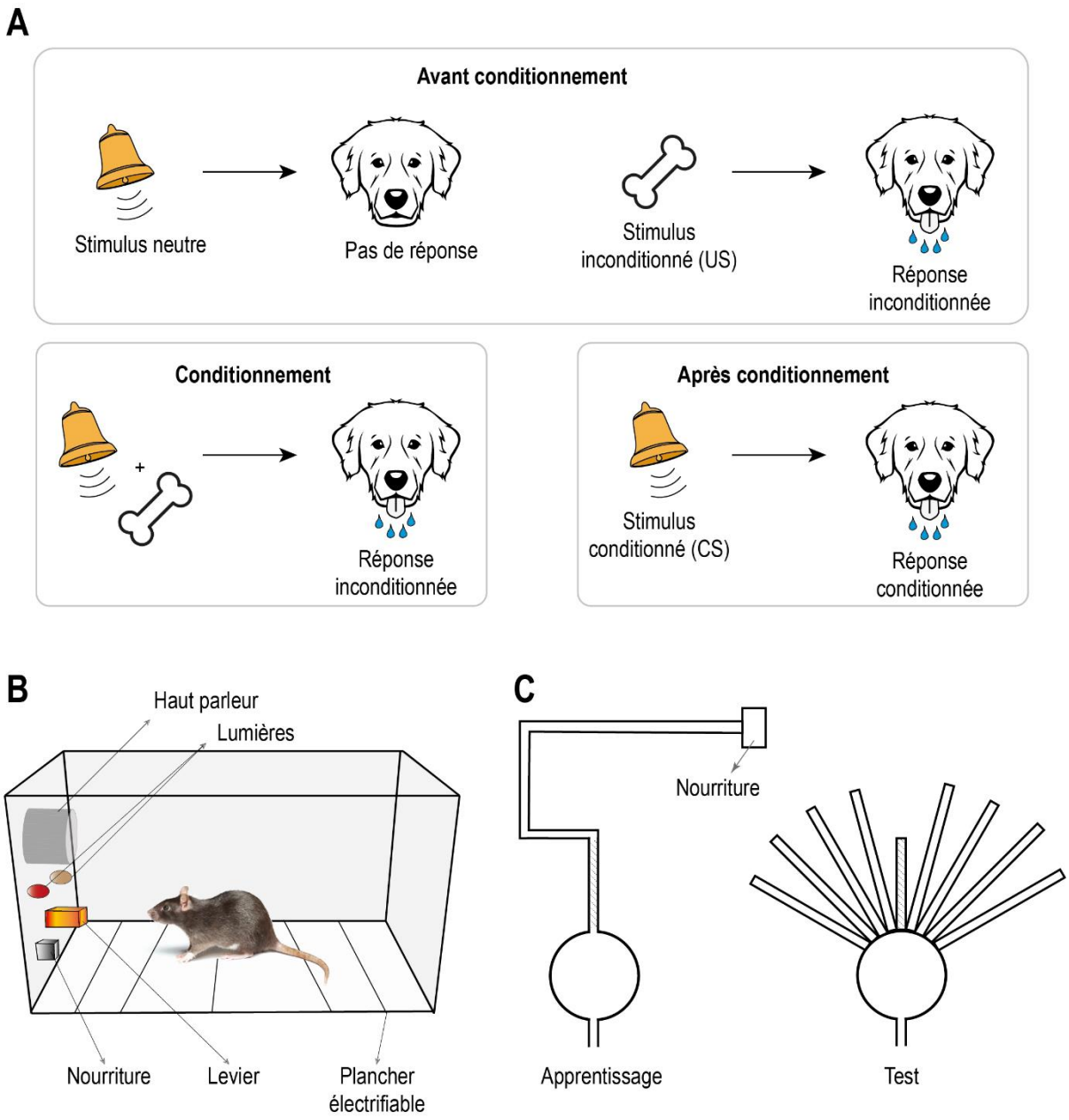


FIGURE 1 : Conditionnement animal

A. Expérience de conditionnement simple de Pavlov. Avant conditionnement, la présentation d'un stimulus neutre n'évoque aucune réponse comportementale tandis que celle d'un stimulus inconditionné (US) évoque une salivation (réponse inconditionnée). Pendant le conditionnement, le stimulus neutre est associé à l'US. Après conditionnement, la présentation du stimulus conditionné (CS) suffit à évoquer la réponse de salivation (réponse conditionnée).

B. Schéma de cage opérante de Skinner.

C. Expérience du raccourci de Tolman. Dans le labyrinthe de conditionnement (gauche), l'animal apprend un chemin menant à la nourriture. Après plusieurs jours d'apprentissage, il est placé dans le labyrinthe de test dont la zone de départ est identique au labyrinthe précédent et dont le bras central connu est bloqué. Le test prend fin lorsque l'animal choisit l'un des bras radiaux.

A cette définition font écho les travaux d'Edward Thorndike[2], psychologue américain, qui s'intéresse chez l'animal au conditionnement dit opérant, souvent opposé au conditionnement simple car reposant quant à lui sur une association entre une réponse instrumentale donc active (séquence d'actions) et un stimulus. Thorndike formule en 1911 la loi de l'effet qui propose qu'une réponse comportementale est davantage susceptible d'être reproduite si elle mène à une récompense satisfaisante (Beeler, 2012; Pessiglione, 2014; Schultz, 2006). Ainsi, si la récompense augmente, la fréquence du comportement amenant à cette récompense augmente aussi. C'est cette idée qui influence notablement John Watson lorsqu'il établit les préceptes du courant behavioriste, et qui mène par la suite aux théories d'apprentissage par renforcement et de conditionnement instrumental. Selon les behavioristes, une récompense renforce une association stimulus-réponse : c'est un renforçateur positif. Ils ont beaucoup décrit les procédés de conditionnement et distinguent deux types de stimuli pouvant induire des apprentissages (Schultz, 2015) :

- les stimuli primaires qui sont ceux associés à des récompenses innées, comme la nourriture par exemple, aussi appelés stimuli inconditionnés (US)
- les stimuli secondaires qui acquièrent une valence de récompense suite à un apprentissage associatif simple ou opérant, aussi appelés stimuli conditionnés (CS).

Burrhus Skinner[3] développe au début du 20ème siècle l'étude du conditionnement opérant. Reprenant les théories de Thorndike et Pavlov, il développe un système expérimental appelé « boîte de Skinner » permettant d'étudier de manière automatisée le conditionnement opérant chez le rat (Pessiglione, 2014; Schultz, 2006). Ce matériel (Figure 1B) est constitué d'une boîte possédant des leviers permettant de délivrer de la nourriture, d'indices visuels et sonores (lumières et haut-parleurs) ainsi que d'un plancher pouvant être électrifié. Grâce à cette boîte, Skinner démontre qu'une récompense (ici la nourriture) peut induire un apprentissage : les rats apprennent à appuyer sur le levier pour obtenir leur nourriture. Avec ce dispositif, il est également possible de conditionner le comportement à la présentation d'un ou de plusieurs stimuli sonores ou visuels : par exemple, le rat apprend qu'il doit appuyer sur le levier

[2] Edward Lee Thorndike (1874-1949) est un psychologue américain précurseur du behaviorisme. Ses travaux portent sur l'intelligence animale, desquels il tire des lois d'apprentissage telles que la loi de l'effet. Il travaille également chez l'Homme sur la psychologie de l'éducation.

[3] Burrhus Frederic Skinner (1904-1990) est un psychologue et penseur américain. Fortement influencé par les travaux d'Ivan Pavlov et ceux de John Watson (considéré comme le premier behavioriste), il est le fondateur du behaviorisme radical, approche qui consiste à se focaliser sur le comportement observable déterminé par l'environnement. Il invente dans les années 1930 un dispositif expérimental de conditionnement opérant, communément appelé « boîte de Skinner ». Celle-ci a permis et facilité l'étude approfondie des mécanismes de conditionnement.

uniquement lorsque l'indice lumineux s'allume. Skinner étudie ainsi l'apprentissage par essai / erreur en présence de renforçateur (nourriture) ou de punition (plancher électrifié). Il décrit ainsi différents types de conditionnements :

- le renforcement positif se traduisant par l'augmentation de la fréquence du comportement permettant d'obtenir la récompense
- le renforcement négatif se traduisant par l'augmentation de la fréquence du comportement permettant d'éviter la punition
- la punition positive se traduisant par la diminution de la fréquence du comportement induisant la punition
- la punition négative se traduisant par la diminution de la fréquence du comportement induisant le retrait de la récompense

Ces expériences de conditionnement opérant ont donc introduit la notion de valence (positive ou négative) d'un stimulus. Mais, bien que les psychologues behavioristes distinguent les renforçateurs positifs et négatifs, ils réfutent toute notion de représentation cognitive et considèrent tout comportement observable comme une simple sortie en réponse à une entrée que constitue le stimulus. Suivant une classification proche des types de stimuli induisant des apprentissages (US et CS cités précédemment), deux classes de récompenses ont ainsi été définies : i) les récompenses primaires qui désignent des récompenses ayant intrinsèquement une valence positive (comme la nourriture ou l'eau), et ii) les récompenses secondaires comme par exemple l'argent ou les récompenses émotionnelles (rire, reconnaissance, etc.). La valence est donc en partie innée, mais peut également être acquise sans avoir recours à des représentations mentales.

Néanmoins, certains paradigmes comportementaux mettent en évidence des contre-exemples de cette thèse behavioriste. C'est notamment le cas de la tâche du raccourci d'Edward Tolman[4] (Tolman et al., 1946) reprise depuis de différentes manières (Alvernhe et al., 2011). Dans cette tâche (Figure 1C), des rats non soumis à une quelconque restriction alimentaire sont placés dans un labyrinthe de conditionnement (ou pré-test) dans lequel ils doivent trouver une pellette de nourriture au bout du bras. Ils effectuent 3 essais par jour durant 4 jours et sont testés une unique fois le 5ème jour dans le labyrinthe de test. Ce dernier présente le même chemin de départ mais, à la différence du premier, dispose de plusieurs autres bras. Le bras similaire au

[4] Edward Tolman (1886-1956) est un psychologue américain qui s'est intéressé aux mécanismes d'apprentissage. Bien que se revendiquant comme behavioriste, il soutient que les comportements ne peuvent pas être réduits à des schémas « stimulus-réponse », mais que l'Homme et l'animal agissent selon un « but poursuivi ». Il développe le concept de carte cognitive.

labyrinthe de conditionnement y est bloqué, mais la pellette de nourriture se situe au même endroit. Tolman observe que la grande majorité des rats choisissent du premier coup d'emprunter un des bras permettant de raccourcir leur chemin. Il développe suite à ces observations l'hypothèse que les animaux utilisent des cartes cognitives pour se représenter leur environnement. Un autre exemple suggérant l'existence de représentations mentales est celui de l'apprentissage latent (Spence and Lippitt, 1946). Les auteurs habituent des rats non soumis à une quelconque restriction alimentaire à un labyrinthe en Y dont les deux bras mènent l'un à de l'eau et l'autre à de la nourriture. Après habituation, les animaux sont restreints en eau ou en nourriture et sont replacés dans le même environnement. Les rats restreints en eau choisissent directement le bras menant à l'eau, et ceux restreints en nourriture le bras menant à l'accès aux pellettes de nourriture, sans effectuer d'erreurs. Cette expérience indique qu'un apprentissage est possible sans qu'il soit pour autant observable au niveau comportemental, l'information apprise pouvant être utilisée plus tard dans un autre contexte environnemental ou lorsque l'animal se trouve dans un état interne différent.

Trois facteurs importants pour le conditionnement, et a fortiori pour tout type d'apprentissage supposent des représentations mentales de valence de stimuli ou actions liées à l'obtention de récompense (Schultz, 2006) : la contingence, l'erreur de prédiction et la contiguïté.

- La récompense doit être **contingente** à l'action c'est-à-dire qu'elle doit être obtenue plus fréquemment en présence du stimulus ou de l'action qu'en son absence. C'est par ce biais qu'il est possible d'inférer que l'animal exerce un contrôle instrumental actif quant à l'obtention de récompenses. Par exemple, un rat ayant associé l'appui sur un levier à l'obtention de récompense sera sensible à une dévaluation (la récompense n'est plus délivrée) ou modification de contingence (appuyer deux fois au lieu d'une pour obtenir la récompense). Cela suggère que l'action est associée à une représentation de ce qu'elle permet d'obtenir.

- L'existence d'une **erreur de prédiction** permet d'ajuster la force de l'association. Tant que la prédiction ne reflète pas correctement le lien entre le stimulus (ou l'action) et la récompense, l'erreur de prédiction est non nulle et l'association peut être renforcée (ou dévaluée). Robert Rescorla et Allan Wagner[5] dans leur modélisation mathématique du conditionnement pavlovien en 1972 ont proposé une règle de correction d'erreur permettant

[5] Robert Rescorla (1940-2020) et Allan Wagner (1934-2018) sont des psychologues et théoriciens américains dont les travaux ont porté sur l'apprentissage associatif. Ils sont à l'origine du modèle dit Rescorla-Wagner conceptualisant le conditionnement classique pavlovien. Ce modèle a servi de base au développement de futurs modèles d'apprentissage par renforcement.

d'ajuster la valeur des actions essai par essai. Ils introduisent ainsi le principe d'un calcul d'erreur de prédiction comme la différence entre la récompense obtenue une fois l'action effectuée, et la récompense prédite. Si cette erreur est nulle, la prédiction correspond à la réalité et la valeur de l'action menant à la récompense est appropriée. Si cette erreur est différente de zéro, la valeur de l'action doit être ajustée car elle ne correspond pas à la réalité. Ainsi une erreur positive indique que la récompense est « meilleure » ou « plus grande » que prédite, et renforce le comportement qui y mène. Cela est par exemple le cas avant apprentissage, lorsque la récompense est obtenue de manière inattendue. En revanche, une erreur négative indique que la récompense est « moins bien » ou « moins importante » que prévu, et le comportement qui y mène est alors dévalué. L'erreur de prédiction est alors considéré comme un substrat computationnel permettant d'orienter vers une réduction de l'association dans le cas où stimulus ne prédit plus la récompense (Niv, 2009; Sutton and Barto, 1998).

- Enfin, la récompense doit suivre la présentation du stimulus ou la réalisation de l'action d'un intervalle de temps optimal : sans cette **contiguïté**, l'association ne pourrait se faire. Cela renvoie directement au problème d'assignement de crédit temporel qui consiste à déterminer, dans un environnement complexe, quelle action parmi toutes celles réalisées par l'animal est responsable du déclenchement d'une récompense. Ce problème est d'autant plus compliqué à résoudre que le délai entre le stimulus (ou l'action) et la récompense est long. Pour accorder du crédit correctement à une action ou un stimulus, il faut être capable de voir au-delà des conséquences immédiates et de les intégrer dans le temps comme une succession d'événements contigus.

Le modèle de Rescorla-Wagner explique particulièrement bien l'**effet de blocage**, observé dans un certain paradigme de conditionnement pavlovien.

$A \rightarrow \text{récompense}$

$A + B \rightarrow \text{récompense}$

$B \rightarrow \text{réponse bloquée}$

Si l'animal apprend que deux stimuli A et B (par exemple un son et une lumière) présentés conjointement prédisent l'obtention d'une récompense, mais qu'il a précédemment associé le CS A seul avec cette même récompense, alors lorsque le stimulus B est présenté seul, la réponse de l'animal est bloquée ou fortement diminuée (Steinberg et al., 2013). La prédiction au moment des deux CS conjoints est alors due à la prédiction préalablement apprise liée au stimulus A, et empêche l'association du stimulus B. Ainsi, l'association avec le CS superflu (B) ne peut se faire. Cet effet illustre bien l'importance d'une erreur de prédiction dans

l'apprentissage associatif. Ces prérequis du conditionnement suggèrent que l'intensité et la fréquence du comportement conditionné sont plus que des associations simples avec la valence de la récompense. Le cerveau est capable d'attribuer des valeurs à des stimuli et à des actions qui, bien que neutres au départ, sont affinées (RPE, assignement de crédit) par apprentissage quand une relation de causalité (contingence) existe avec la récompense et est intégrable par l'animal (contigüité). Cette valeur est une prédiction temporelle de ce qui va arriver en fonction d'évènements externes (stimuli) ou internes (actions), et correspond à une forme de représentation mentale, en contradiction avec le dogme béhavioriste.

Le principal apport de l'approche behavioriste dans la compréhension des processus d'apprentissage reste donc le renforcement positif, c'est-à-dire l'idée que la fréquence d'un comportement augmente lorsqu'il est associé à une récompense (renforçateur positif). Si les théories behavioristes ont permis la description fouillée de comportements et notamment du conditionnement, elles ont ignoré les représentations mentales et facteurs internes pouvant influencer les comportements tels que les connaissances au préalable, ou la notion d'état. Néanmoins, le raffinement des protocoles de conditionnement par les behavioristes a introduit la notion de valence et conduit à celle de valeur, notion à laquelle les mouvements cognitivistes s'intéresseront dans la deuxième moitié du 20ème siècle.

3) Récompense et motivation vers un but

Le mouvement cognitiviste postule que si l'intensité d'un comportement tourné vers une récompense est modulée par la quantité de récompense, c'est que l'animal se représente la valeur de ses actions et des récompenses associées. En effet, outre leur effet hédonique immédiat, les récompenses suscitent également à long terme de la motivation qui pousse les individus à les obtenir. Un individu motivé par une récompense peut travailler et est prêt à fournir un effort pour l'acquérir (Salamone and Correa, 2012). La récompense constitue alors un but à atteindre. Le mouvement cognitiviste étudie l'aspect de la représentation de buts notamment au travers du concept de motivation. Ce mouvement se détache donc de l'approche fréquentielle adoptée par les behavioristes, et permet d'éclaircir ce que ces derniers expliquent mal, comme l'apprentissage latent. L'existence de buts et d'un *drive* motivationnel permet de mieux appréhender ces observations.

Le terme *motivation* a endossé diverses définitions au cours du XXème siècle, considéré comme un procédé de maintien d'un comportement en cours, ou une force poussant à agir pour obtenir satisfaction (on parle aussi de *drive*). Nous pouvons retenir une définition plus récente qui désigne par *motivation* le procédé par lequel l'organisme oriente son comportement afin de prendre le contrôle sur son environnement externe et interne selon ses besoins (Di Chiara, 1995).

Dès le début du XXème siècle, l'éthologue Wallace Craig décrit la motivation comme un procédé divisé en deux phases successives (Berridge, 2004; Craig, 1917; Salamone and Correa, 2012) :

- une **phase appétitive**, aussi dite anticipatoire, préparatoire (ou plus récemment de recherche), qui consiste en un état d'agitation et un comportement d'approche flexible permettant d'atteindre le but fixé.

- une **phase consommatoire** dans laquelle le comportement de consommation est enclenché par l'obtention du but. Cette phase est en quelque sorte la finalité du processus motivationnel, qui prend fin une fois la récompense consommée.

Une première vision des processus motivationnels dans les années 1960 est qu'ils sont déclenchés par des stimuli saillants. Ces stimuli acquièrent avec l'expérience une valence motivationnelle, suite à un conditionnement par exemple, puis suscitent de la motivation qui peut déclencher le comportement. Mais la motivation ne se limite pas à déployer une réponse instrumentale de manière innée ou conditionnée, ni ne se restreint à l'intensité du comportement de consommation de la nourriture. Le point central de la motivation est qu'elle doit aussi pouvoir orienter la réponse instrumentale de manière la plus flexible possible et permettre de s'adapter à de nouvelles contingences (Berridge, 2004). Partant de ce principe, trois autres aspects sont introduits dans les années 1970-80 permettant de mieux appréhender les mécanismes de motivation :

- la **flexibilité** : un individu doit être capable de modifier la manière d'atteindre son but si les circonstances changent mais que le but demeure appétitif. C'est par exemple le cas lors de l'inférence d'un nouveau chemin pour atteindre un but, si le chemin habituellement emprunté est bloqué (comme pour l'expérience du raccourci, cf. Chapitre 1 – I) – 2)).

- l'**anticipation** du but : cela implique une forme d'apprentissage associatif du but qui est appétitif. L'anticipation peut se mesurer par la vigueur de l'action menant au but. Plus un animal est motivé, plus il se dirigera rapidement vers le levier distribuant de la nourriture par exemple.

- l'**affect** : la motivation dirige un comportement vers un but hédonique, c'est-à-dire provoquant du plaisir. Atteindre le but produit une réaction comportementale ou physiologique traduisant un état émotionnel positif. Notons que la direction du comportement motivé peut être soit d'aller vers une récompense soit de s'éloigner ou d'éviter un stimulus négatif (douleur ou inconfort par exemple).

Ces trois aspects sont les principes de bases de la théorie dite de la **motivation incitatrice** (Berridge, 2004; Bindra, 1974; Bolles, 1972; Di Chiara, 1995; Jensen and Toates, 1993). Développée dans les années 1970, elle propose que la motivation incite à l'action par une anticipation caractérisée par une augmentation d'activité motrice pouvant aller jusqu'au déclenchement en avance du comportement. Selon cette théorie, le caractère hédonique des récompenses permet par apprentissage de percevoir des stimuli prédicteurs (CS) comme des récompenses (US) en soi. Plus qu'une réponse spécifique, les CS initient un état général de motivation dirigée vers les US, état dans lequel l'animal peut moduler ses réponses pour atteindre son but. Cette définition s'oppose donc à l'association rigide entre un stimulus et une action amenant à une récompense. Les comportements de recherche (*wanting*) qui sont déclenchés participent alors à l'obtention de ces buts ou récompenses (Bindra, 1974; Bolles, 1972). On parle de l'effet directionnel de la motivation.

La motivation incitatrice peut être modulée selon l'état interne de l'individu et ses éventuels besoins fondamentaux (faim, soif, période de reproduction, etc.) (Jensen and Toates, 1993). On parle alors d'effet activateur de la motivation (Salamone and Correa, 2012) : cela permet à l'individu d'allouer l'énergie nécessaire pour surmonter les éventuels obstacles physiques ou bien la distance temporelle le séparant de l'obtention du but. La motivation pousse à effectuer un effort malgré d'éventuelles conséquences néfastes (coût moteur, coût temporel, prise de risque pour sa survie, etc.) et induit la dynamisation des actions permettant l'obtention du but, ce qui se traduit notamment par une augmentation de vitesse de déplacement ou d'exécution de l'action. La motivation peut être mesurée assez facilement en laboratoire pour des récompenses primaires en limitant l'accès à la nourriture ou à l'eau et en conditionnant cet accès. La motivation incitatrice permet ainsi d'expliquer les phénomènes d'apprentissage latent et d'inférence de nouveaux chemins spatiaux.

Dans les théories classiques (Berridge, 2004), la motivation incitatrice est induite par un stimulus et est extrinsèque : les comportements sont générés par un stimulus externe pour parvenir à un but également externe. Cependant, les actions générées « spontanément » (en absence de stimulus incitateur) pourraient être motivés par des représentations internes, basées sur la mémoire, des récompenses (Passingham et al., 2010). De même, au cours de la vie d'un

individu, toute action n'est pas tournée vers une récompense physique. Certains comportements sont réalisés pour eux-mêmes (jeu, entraînement, etc.) plutôt que dirigés pour obtenir un résultat particulier et anticipé, on parle alors de motivation intrinsèque. C'est le cas également des comportements de curiosité, d'exploration et de recherche (Oudeyer and Kaplan, 2009).

L'étude de la motivation distingue la performance comportementale (intensité, fréquence d'une action, vitesse d'exécution) de l'apprentissage des contingences entre comportement et environnement (Balleine, 2019), ce qui met à jour l'existence de différentes représentations dans le cerveau (valeurs de récompense, état interne, paramètres de l'environnement, etc.). L'apprentissage ne génère pas nécessairement un changement comportemental (apprentissage latent par exemple), et un changement comportemental peut refléter différents changements internes (apprentissage mais aussi modification d'état interne ou de but). Les influences multiples sur les décisions suggèrent qu'il existe différents systèmes d'attribution de valeur, que je vais maintenant exposer.

II) Différents systèmes d'attribution de valeur

Trois systèmes d'attribution de valeurs sont distingués, selon le type d'association entre comportement et environnement qu'ils engagent : le système pavlovien qui concerne des comportements non volontaires déclenchés par un stimulus, le système des comportements orientés vers un but, qui se basent sur des associations entre une action et son résultat, et le système des comportements habituels qui se basent sur des associations stimulus-réponse (Dayan, 2009; Dolan and Dayan, 2013; Rangel et al., 2008).

1) Le système pavlovien

Le système pavlovien concerne un ensemble de comportements qui sont des réponses automatiques et réflexes à des stimuli particuliers de l'environnement. Il fait référence à l'expérience de conditionnement classique menée par Ivan Pavlov (cf. Chapitre 1 – I) – 2), Figure 1A) en 1890. Ces comportements sont des comportements innés (salivation,

mordillements, fermeture des paupières ou *eye blinking*, immobilisation ou *freezing*, comportements de fuite ou d'agression, etc.) en réponse à des stimuli environnementaux (présence de nourriture, de partenaires, de prédateurs, etc.). Ils peuvent également être acquis par conditionnement, et être déclenchés dans ce cas en réponse à des stimuli prédicteurs (CS), (salivation en réponse au son de la cloche dans l'expérience de Pavlov). Ce dernier, aussi dit « conditionnement pavlovien », est un apprentissage associatif stimulus-récompense (S-US), à différentier du conditionnement instrumental qui repose sur l'apprentissage par renforcement d'une chaîne d'actions (Berridge and Robinson, 2003) (voir plus bas). Les comportements d'approche de récompenses et des stimuli qui les prédisent sont aussi considérés comme des comportements pavloviens (Dayan et al., 2006; Dayan and Berridge, 2014) et constituent une faculté adaptative majeure. Les actions déclenchées par le système pavlovien ne sont en revanche pas flexibles ou motivées par un but (cf. Chapitre 1 – I) – 3)) et ne sont donc pas considérées comme volontaires.

Cependant, le système pavlovien reste sous influence motivationnelle et peut être mis en en jeu dans des apprentissages latents. Dans l'expérience du « sel de la Mer Morte » (Robinson and Berridge, 2013), des rats associent un stimulus à une nourriture très salée, qui n'est pas leur nourriture favorite initialement - le stimulus conditionné n'induit ici pas d'approche. Puis les animaux sont mis dans un état de carence en sodium, et, dès leur première exposition à la tâche dans ce nouvel état physiologique, approchent en premier lieu les stimuli associés aux aliments les plus salés. Les auteurs proposent que l'état de besoin en sel induit une transformation hédonique « instantanée » (non apprise par essai et erreur) des stimuli liés aux aliments salés, ce qui permet à ces derniers d'être appréciés lorsqu'ils sont consommés dans cet état-ci, l'approche du stimulus associé à la salinité n'est pas volontaire (il est automatique en état de privation de sel) mais reste sous influence motivationnelle de l'état interne de l'animal.

Le système pavlovien peut également produire des comportements inadaptés. C'est notamment le cas de l'*autoshaping* qui désigne une forme de conditionnement dans laquelle l'animal associe un stimulus à une réponse qui est systématiquement accompagnée d'un comportement non pertinent car non-contingent à l'obtention de la récompense. Cela a été initialement observé chez le pigeon : dans une tâche dans laquelle une lumière précédait la distribution de nourriture, certains animaux picoraient l'indice lumineux lorsqu'il s'allumait (de la même manière dont ils consommaient la nourriture), alors que cette action n'est pas nécessaire à l'obtention de la nourriture (Dayan et al., 2006). Les inconvénients et le caractère inadapté pouvant être générés par le système pavlovien sont potentiellement mis en jeu dans

des dérèglements comportementaux comme les stéréotypies, les troubles obsessionnels compulsifs (TOCs) ou les phénomènes de rechute dans l'addiction (Rangel et al., 2008).

2) Le système orienté vers un but

Les comportements orientés vers un but sont entièrement basés sur des associations action-résultat. Autrement dit, c'est la récompense constituant le but qui motive le comportement. Contrairement aux comportements pavloviens, les comportements orientés vers un but ne sont pas innés, mais composés de chaînes d'actions volontaires constituant le moyen d'atteindre le but. On parle de comportement instrumental (Berridge and Robinson, 2003).

Le système orienté vers un but ajuste les valeurs des associations action-résultat dès que la valeur du résultat change, ce qui permet une flexibilité de ces comportements dans des environnements dynamiques ou instables. Ces comportements reposent donc sur la connaissance et la représentation à la fois de l'état de l'environnement, de l'état de l'individu, de sa motivation, et du but à atteindre. La vision computationnelle de ces comportements les modélise comme des procédés *model-based* (Figure 2), c'est-à-dire qu'ils sont basés sur un modèle cognitif de la situation, que l'on peut représenter sous la forme d'un arbre des états, actions et résultats possibles (Dayan, 2009; Dayan and Niv, 2008). Dans ce cadre théorique, l'erreur de prédiction de récompense sert à apprendre les transitions entre des états de l'environnement qui sont déclenchés par une action de l'animal (par exemple : appuyer sur un levier permet de l'abaisser, tourner à gauche aboutit à telle partie du labyrinthe). Elle permet aussi d'apprendre les conséquences de ces transitions, autrement dit la valeur de l'état final (par exemple : abaisser *n* fois le levier, ou arriver à tel endroit du labyrinthe amène une récompense). Cet apprentissage est flexible car les transitions et les conséquences sont mémorisées indépendamment. Ainsi, si les transitions sont modifiées (comme la tâche du raccourci dans le labyrinthe de Tolman, cf. Chapitre 1 – I) – 2)) ou si la valeur des états change (comme dans l'apprentissage latent et les changements d'états motivationnels), la valeur des actions est recalculée. Ainsi, une fois ce modèle construit, l'individu peut « se déplacer » dans cet arbre à la recherche d'une action appropriée pour parvenir à son but et en sélectionner une.

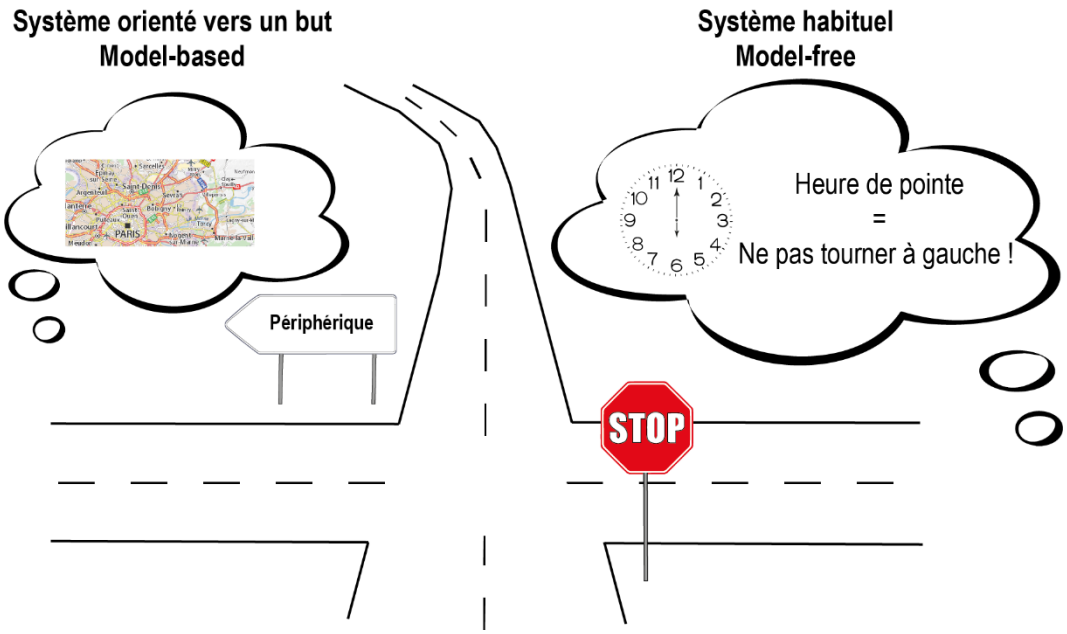


FIGURE 2 : Deux systèmes de comportements volontaires

Représentation schématique de deux manières de se comporter à une intersection, selon que le système qui prend le contrôle est le système orienté vers un but ou le système habituel. *Adapté de Dayan & Niv, 2008*

Ce mode de fonctionnement est très exigeant et couteux d'un point de vue computationnel car nécessite des représentations mentales multiples de l'environnement ainsi que de l'état interne, et notamment du niveau motivationnel (Dolan and Dayan, 2013; Miller et al., 2018; Schultz, 2015). Selon cette approche, afin de mettre à jour ses représentations et les valeurs attribuées aux actions, le cerveau doit en effet évaluer a posteriori la désirabilité ou l'utilité des actions réalisées. Par ailleurs, c'est un système dit délibératif qui laisse place à de possibles conflits entre différents buts, ce qui dans ce cas revient à choisir un but parmi plusieurs (Rangel et al., 2008). Par exemple, on peut désirer manger une glace en été, en suivant un premier but à court-terme qui est de ressentir le plaisir du goût et de la fraicheur de la glace ; mais cela entre en conflit avec un second but à long terme qui est de limiter ses apports en ne pas manger trop de sucre pour préserver sa santé et sa ligne. On considère donc que les comportements orientés vers un but ne sont pas le fruit d'un simple conditionnement associatif déclenché par un stimulus, mais sont des comportements volontaires instrumentaux qui nécessitent une représentation mentale du but (ou de multiples buts) et des actions nécessaires pour l'atteindre.

Expérimentalement, pour tester si un comportement est orienté vers un but, on procède généralement à une dévaluation ou à une dégradation de contingence (Dolan and Dayan, 2013). Par exemple, pour un rat ayant appris à appuyer sur un levier pour obtenir sa nourriture, la

dévaluation consistera simplement à ne plus associer l'appui du levier avec la distribution de nourriture, le levier devenant dans ce cas un objet neutre. Ces protocoles aboutissent à une extinction du comportement orienté, extinction qu'on ne retrouve pas de manière similaire dans les comportements de type habituel.

3) Le système habituel

Souvent opposés aux comportements orientés vers un but, les comportements habituels sont appris par l'expérience et la répétition. Par essai-erreur, un individu attribue une valeur à une association stimulus-réponse comportementale, cette valeur est ajustée à chaque répétition et correspond in fine à la récompense attendue résultant du comportement. Ce système se caractérise par un apprentissage plutôt lent et très résistant à l'extinction. Après un long entrainement, le contexte décisionnel seul déclencherait le comportement, sans que le but ait besoin d'être mentalement représenté. Les comportements habituels sont, une fois appris, rapidement enclenchés et peu flexibles (Dolan and Dayan, 2013; Rangel et al., 2008; Schultz, 2015).

L'apprentissage des comportements habituels a été largement conceptualisé par la théorie de l'apprentissage par renforcement (Niv, 2009; Sutton and Barto, 1998) dont la notion clé est la maximisation des récompenses. Dans ce cadre, la valeur d'une action est simplement incrémentée de l'erreur de prédiction. Les valeurs des actions retenues sont celles des essais passés, et sont dissociées de toute prédiction des futures conséquences des actions. Elles ne sont donc pas recalculées en cas de changement dans l'environnement. D'un point de vue computationnel, cette stratégie d'apprentissage est donc dite « sans modèle » (*model free*, Figure 2) (Dayan, 2009; Dolan and Dayan, 2013; Kakade and Dayan, 2002), par opposition à la stratégie « basée sur un modèle » (*model based*) (cf. Chapitre 2 – II) – 2)) : il n'y a pas d'apprentissage de l'effet de l'action sur l'environnement, mais seulement de sa valeur (par exemple : appuyer sur un levier amène de la nourriture, tourner à gauche est récompensé, etc.). La répétition et l'expérience ont permis de ne retenir qu'une valeur unique d'association stimulus-réponse, et l'individu n'a plus besoin de rechercher dans l'arbre des actions possibles l'action la plus adaptée. Les comportements habituels sont assez résistants aux protocoles d'extinction et en général longs à désapprendre. Ils sont donc potentiellement peu efficaces si

l'individu évolue dans un environnement dynamique avec des probabilités élevées de changement.

D'autres auteurs proposent un découpage différent dans les stratégies d'apprentissages. Dans ce cadre, les comportements dirigés vers un but sont directement orientés par la valeur de l'action (*value based*), ce qui explique leur degré de flexibilité. Un comportement habituel est au contraire réalisé uniquement en fonction du nombre de fois où il a été expérimenté dans le passé, et est dans ce cas « sans valeur » (*value free*) (Miller et al., 2019, 2018).

On considère en général qu'un comportement est d'abord orienté vers un but, puis devient habituel dans un second temps. En effet, dans les deux cas, c'est une réponse instrumentale qui est renforcée. Avec les répétitions, la représentation des conséquences des actions devient moins forte que la valeur de l'association stimulus-réponse qui constitue l'habitude. Cela suggère que les différents systèmes d'attribution de valeur ne sont pas séparés et interagissent.

4) Interactions entre systèmes d'attribution de valeur

Ces différents systèmes d'attribution de valeurs ont été distingués à la fois sur le plan expérimental et computationnel, mais ils ne sont pas indépendants les uns des autres. Il est possible par exemple que le système pavlovien soit mis en jeu dans un premier temps avant le système orienté par les buts, et que le système habituel entre en jeu seulement en dernier par répétition du comportement (Rangel et al., 2008).

Certains travaux décrivent par exemple des phénomènes de transfert pavlovien-instrumental. Ce transfert peut être observé dans certaines conditions lorsqu'un stimulus conditionné facilite ou augmente l'intensité d'une réponse instrumentale (Estes, 1943). Dans ces expériences, un rat dans une cage opérante apprend d'abord à appuyer sur un levier pour obtenir sa nourriture. Dans un second temps et dans un autre environnement, il associe la distribution de nourriture avec un stimulus conditionné, une lumière par exemple. Enfin, replacé dans la cage opérante, le rat appuiera davantage sur le levier lorsque l'indice lumineux sera présenté. Ce type d'expériences indique explicitement que des stimuli saillants peuvent moduler la valeur motivationnelle et incitatrice d'un comportement opérant (Bogacz, 2020; Dickinson and Balleine, 1994). Le transfert pavlovien-instrumental pourrait être un des

mécanismes susceptibles de provoquer des rechutes ou de la surconsommation dans le cas d'addiction à des drogues d'abus.

Certains auteurs suggèrent que le système à qui est donné le contrôle du comportement est celui qui estime les valeurs des actions de la manière la plus certaine (Daw et al., 2005; Rangel et al., 2008). Puisque le système habituel attribue des valeurs avec de plus en plus d'acuité au cours du temps, il devrait théoriquement prendre plus souvent le contrôle sur le système orienté vers un but, permettant ainsi « d'économiser » des ressources cognitives pour des problèmes plus exigeants à résoudre.

Il existe également des conflits entre ces trois systèmes d'apprentissage (Dayan et al., 2006) qui seraient notamment déterminants dans les perturbations comportementales causées par des pathologies telles que l'addiction (Rangel et al., 2008). Considérons un individu alcoolique qui hésite à prendre une bière en rentrant chez lui après sa journée de travail. Le système habituel accordera une forte valeur à la bière car le comportement tourné vers la consommation de l'alcool a été largement renforcé par l'expérience passée. En revanche, le système orienté vers le but d'évitement de consommation d'alcool accordera une faible valeur à la bière pour diminuer l'incitation à la consommer.

L'existence de tels conflits entre différentes méthodes d'apprentissage de valeurs indique bien que l'aspect récompensant d'un comportement ou d'un objet peut se situer à différents niveaux. Une récompense peut induire un apprentissage de différentes manières et constituer un but à court ou long terme, ce qui fait d'elle une notion clé de la prise de décision.

III) Prendre une décision

Les animaux, dont l'Homme, sont capables d'associer des valeurs à des récompenses et disposent de multiples représentations internes leur permettant d'orienter leur comportement. Il existe divers procédés de sélection pour aboutir à des comportements appropriés et les contrôler. Face à un choix entre plusieurs alternatives, comment un individu choisit-il ? Dans ces processus de prise de décision, la récompense est bien évidemment un objet crucial puisqu'elle constitue la conséquence d'un choix ou d'une action et suscite de la motivation. Choisir entre plusieurs options revient en quelque sorte à choisir entre plusieurs récompenses. Mais ce n'est pas toujours la plus grande récompense qui est choisie. En effet, lorsqu'un

individu prend une décision, il se réfère également aux informations à la fois sur son état interne, le contexte dans lequel il se trouve, son but, l'effort nécessaire à fournir pour planifier et exécuter l'action, et les critères selon lesquels il sélectionnera un plan d'action approprié.

1) Choix économique

Les théories de prise de décision reposent sur la notion de choix économique qui est défini comme le comportement résultant d'un choix basé sur des préférences subjectives (Padoa-Schioppa, 2011). Étant donné que le temps est limité et qu'un individu ne peut réaliser qu'une action à la fois, le principe d'un choix économique est de considérer qu'il y a une compétition entre différentes représentations afin qu'une d'entre elle déclenche le comportement de choix. Ce choix découlerait donc d'une comparaison des valeurs associées à chaque issue possible. Cette notion centrale en théorie économique depuis le 17ème siècle ainsi qu'en psychologie expérimentale, est un centre d'intérêt majeur des neurosciences depuis ces dernières décennies. L'un des principes de la théorie économique, est qu'un individu choisit de manière optimale et rationnelle l'option ayant la plus forte valeur, et ce afin de maximiser son utilité (Edwards, 1954; Glimcher, 2009; Padoa-Schioppa, 2011). La valeur subjective d'une récompense correspond à son utilité, qui reflète en quelque sorte sa désirabilité selon l'état interne dans lequel le sujet se trouve. Dans le cadre de cette théorie, l'un des principes fondamentaux est la transitivité des préférences. Si un individu préfère une tarte au citron à une tarte aux pommes, et une tarte aux pommes à un gâteau au chocolat, alors il doit également préférer la tarte au citron au gâteau au chocolat. Cette règle de transitivité (si A>B et B>C alors A>C) doit être invariante et indépendante du contexte (Edwards, 1954). En pratique, de nombreux exemples de choix irrationnels s'opposant à ce principe ont été relevés depuis lors : il est notamment admis que la manière dont un choix est proposé (l'effet de cadrage, ou *framing effect*), ou encore l'état interne et l'histoire personnelle de chaque individu influencent et biaisent la prise de décision. Néanmoins, ce schéma normatif a constitué un cadre de réflexion aux études neuroéconomiques.

La neuroéconomie est l'étude des bases neurobiologiques et computationnelles de la prise de décision basée sur la valeur. Son objectif est de fournir une description algorithmique des mécanismes de choix chez l'Humain. Les modèles de prise de décision proposent généralement que ce soit un processus en deux étapes. La première consiste en l'évaluation des options

possibles. La seconde étape est le choix d'une option parmi l'ensemble des celles qui sont présentées. Ce choix est biaisé de manière stochastique vers l'option de plus haute valeur ou utilité (Glimcher, 2009; Padoa-Schioppa, 2011). Un point important de cette modélisation est que le processus neuronal de choix intervient en amont de la planification et de l'exécution de l'action associée à ce choix.

Le cadre d'étude le plus simple de la prise de décision en neuroéconomie est celui du choix binaire. Choisir entre A ou B est un problème de décision permettant de modéliser les processus de comparaison de valeur et les stratégies de décisions. Il est largement utilisé aussi bien chez l'Homme que chez l'animal.

2) Valeur et utilité

La plupart des théories de prise de décision reposent sur l'idée que les valeurs des actions possibles sont représentées et confrontées au moment du choix. Certains formalismes proposent des représentations sous une forme absolue, ou bien sous formes de valeurs dépendantes du contexte, tandis que d'autres proposent des représentations uniquement basées sur des comparaisons entre options, c'est-à-dire des valeurs relatives (Vlaev et al., 2011). L'idée dominante est que lorsque les valeurs sont comparées, l'option ayant la plus grande valeur subjective, appelée dans ce cadre utilité, a tendance à l'emporter.

L'attribution d'une valeur à une option permet d'intégrer sous forme d'une échelle commune divers déterminants associés à cette option, lesquels varient parfois sur des dimensions différentes (Glimcher, 2009; Schultz, 2015, 2006). Les valeurs subjectives des différentes options peuvent ainsi être confrontées. Par exemple, lorsqu'il s'agit de choisir entre une maison plus grande mais plus loin du lieu de travail, ou une maison plus proche mais plus petite, les éléments qui entrent en compte dans la décision ne sont pas du même ordre ; il faut pourtant pouvoir mettre les deux options face à face (sur une échelle commune) pour en sélectionner une. Que la valeur soit attribuée par le système habituel ou le système orienté vers un but, elle est modulée par différents paramètres tels que :

- la **nature** et la **qualité de la récompense**. Ces critères influencent la valeur selon des préférences personnelles. Un individu peut préférer le goût de la pomme à celui de l'orange et lui attribuer de manière générale une plus grande valeur.

- la **quantité de récompense** obtenue à l'issue du choix : une grande récompense aura généralement une grande valeur.

- le **coût** associé à l'action menant à la récompense. Il peut être énergétique, autrement dit moteur (par exemple faire un long trajet pour aller au restaurant), ou bien cognitif (par exemple organiser un voyage qui satisfera entièrement mes attentes), et fait diminuer la valeur de l'option. Pour une même quantité de récompense, si une option présente un coût plus faible qu'une autre, elle aura une plus grande valeur.

- le **délai temporel** avec lequel est obtenue la récompense : les actions menant à des récompenses trop décalées dans le temps sont associées à des valeurs plus faibles, on parle de décomptage temporel (*delay discounting*) (Rangel et al., 2008).

- l'**incertitude** quant à l'obtention de la récompense : un faible degré de certitude sur l'issue d'une décision complique son estimation et a tendance à abaisser la valeur. Dans un environnement naturel, les conséquences des actions qu'entreprennent les animaux sont souvent probabilistes, il est donc très fréquent qu'il y ait une part d'incertitude dans nos estimations. Deux types d'incertitude sont distingués, l'incertitude attendue et inattendue (Heilbronner et al., 2010; Yu and Dayan, 2005). L'incertitude attendue reflète la connaissance d'une relation incertaine entre une action et son résultat, peut de ce fait être approximée, et est considérée comme incompressible. Au contraire, l'incertitude inattendue est provoquée par des changements de contingences dans l'environnement dont la probabilité de survenir ne peut pas être connue. Lorsque l'on choisit un restaurant sur la base de divers avis croisés de différentes sources, on tente d'estimer au mieux une forme attendue d'incertitude. En revanche, il est possible qu'une fois rendu au restaurant choisi, les propriétaires et l'équipe en cuisine aient changé depuis la veille. Cette forme d'incertitude inattendue est difficilement prévisible. De la même manière, le **risque** est un paramètre influençant la valeur d'une récompense. A la différence de l'incertitude, le risque désigne la probabilité que l'action choisie soit associée à une conséquence de valence négative, comme une punition ou une perte. Un piéton qui courre au milieu d'une intersection où circulent des voitures arrivera peut-être plus vite à destination mais prend le risque de se blesser.

- la **satiété** ou **l'ennui** (*boredom*), impliqués notamment lors de la prise répétée de récompenses de même nature, ont tendance à diminuer la valeur subjective de la récompense suivante. A l'inverse, la **nouveauté** peut représenter un attrait particulier pour une option jamais rencontrée auparavant (cf. Chapitre 4 – IV) – 2)) (Wilson et al., 2014a).

- le degré de **motivation**, le danger potentiel, l'ambiguïté, le contexte social, etc. sont d'autres paramètres influençant la valeur associée à une option.

Certains de ces déterminants peuvent être appris ou estimés par expérience (comme le risque d'attraper un coup de soleil après avoir déjeuné en plein soleil), d'autres sont liés à des croyances a priori (un vin plus cher aura une plus grande valeur subjective même s'il n'est pas forcément qualitativement meilleur) ou à l'état interne (tel que le niveau de satiété au moment du choix).

Certains comportements ont, par ailleurs, une valeur intrinsèque, qui ne dépend pas du résultat de l'action, on parle aussi de récompense intrinsèque (Oudeyer and Kaplan, 2009; Schultz, 2015). C'est le cas par exemple des comportements d'exploration ou de jeu, beaucoup observés chez les enfants, qui permettent d'accéder à des récompenses d'ordre supérieur (par exemple jouer d'un instrument de musique ou pratiquer une activité sportive). Plusieurs paramètres influencent la valeur d'une récompense intrinsèque : la nouveauté, l'imprévisibilité, la sensation de s'améliorer (Oudeyer and Kaplan, 2009). Le coût moteur de certains mouvements ou efforts peut être complètement contrebalancé par leur aspect intrinsèquement plaisant. Par ailleurs, les récompenses intrinsèques étant souvent associées à des récompenses extrinsèques (par exemple, explorer une nouvelle parcelle de terrain dans l'espoir de trouver une source de nourriture plus abondante), l'utilité nette qui s'en dégage comprend à la fois une composante intrinsèque, et une liée à la valeur extrinsèque du résultat (Schultz, 2015).

Face à une offre, un individu choisit en tenant compte de nombreux paramètres qui biaisent sa décision selon plusieurs dimensions. En particulier, une tendance à choisir l'option ayant la plus haute espérance se dégage. Ce principe de maximisation de l'utilité a été formulé par Blaise Pascal au milieu du 17ème siècle : un individu évalue les conséquences des différentes options possibles et choisit celle qui a le plus haut gain attendu (Pessiglione, 2014; Schultz, 2006). Plusieurs manières d'intégrer l'ensemble de ces influences ont été proposées. La théorie de l'utilité attendue, basée sur les travaux de Daniel Bernoulli au 18ème siècle et développée par John von Neumann and Oskar Morgenstern à la fin du 20ème siècle propose que la valeur subjective d'une option soit pondérée par sa probabilité d'être récompensée (Edwards, 1954). Ainsi, après un nombre suffisamment important de choix, la distribution des choix reflètera la valeur attendue de chaque option. La fonction d'utilité ne croit pas linéairement avec le gain, mais est généralement concave (Glimcher, 2009; Padoa-Schioppa, 2011; Rangel et al., 2008). Autrement dit, dans la plupart des situations, gagner 200€ n'est pas deux fois plus utile que de gagner 100€.

Ce formalisme permet également de décrire certains comportements comme l'aversion au risque ou la recherche de risque (Figure 3). D'un individu à un autre, la fonction d'utilité

peut être plus ou moins concave voire convexe (recherche du risque). Toutes ces théories sous-entendent qu'il existe dans le cerveau des représentations des valeurs subjectives des actions. Afin d'être les plus fiables possibles, ces représentations doivent être continuellement mises à jour en fonction du résultat d'un choix : l'utilité d'une issue est réévaluée a posteriori afin d'apporter un retour d'information.

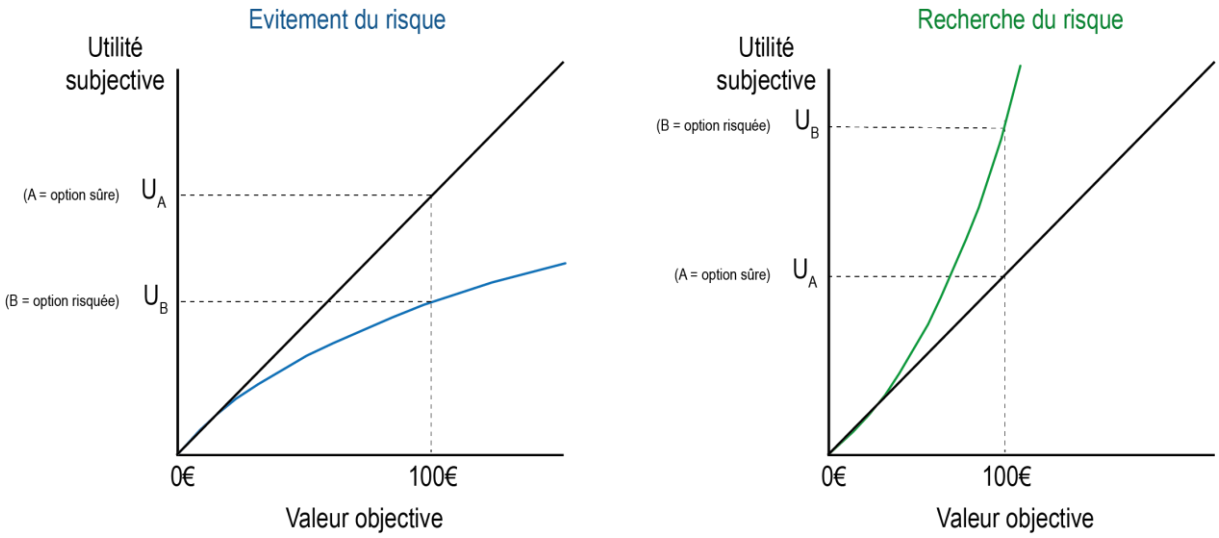


FIGURE 3 : Fonctions d'utilité et attitude face au risque

Représentations graphiques de l'utilité subjective d'une option en fonction de sa valeur objective. L'unité monétaire en € a été choisie pour l'exemple. La situation est la suivante : un individu a le choix entre gagner 100€ à coup sûr (option A) ou 200€ avec une probabilité de 50% (option B). La valeur objective (espérance mathématique) des deux options est de 100€. Les droites noires représentent une relation linéaire entre la valeur objective et l'utilité subjective. Gauche : courbe d'utilité concave d'un individu aversif au risque. $U_A > U_B$: l'individu préfère le choix sûr A. Droite : courbe d'utilité convexe d'un individu recherchant le risque. $U_B > U_A$: l'individu préfère le choix risqué B. Bien que la valeur objective de gain soit la même pour les deux options, le premier individu (gauche) préfère un gain plus faible mais sûr, tandis que le second (droite) prend le risque de parier en espérant obtenir un plus gros gain.

D'autres théories plus complexes, comme la théorie des perspectives (*prospect theory*) développée par Daniel Kahneman et Amos Tversky en 1979, définissent une fonction d'utilité non linéaire qui explique bien les différents biais de décision observés chez l'animal dont l'Homme. Par exemple, de nombreuses personnes jouent chaque jour au loto en surévaluant des probabilités de gain infimes, tout en négligeant le risque de se faire frapper par la foudre (tout aussi infime soit-il). Les individus évaluent de manière asymétrique leurs perspectives de perte ou de gain, et cela dépend d'une référence interne et subjective (Constantinople et al., 2019; Rangel et al., 2008).

3) Politique

Sur la base des utilités des différentes options, un individu peut mettre en œuvre différentes stratégies (ou politiques) de prise de décision qui guideront son comportement de différentes manières. D'un individu à un autre, ou bien selon le contexte, les stratégies de choix peuvent être plus ou moins orientées vers les options de plus haute valeur, plus ou moins aléatoires ou variables. La capacité à mettre en œuvre diverses stratégies de décision permet d'assurer la richesse du comportement.

La politique de choix visant à maximiser systématiquement la récompense obtenue est qualifiée d'exploitation. Au contraire, les politiques intégrant plus de variabilité dans les décisions sont des stratégies dites d'exploration. Exploiter les ressources connues et disponibles dans son environnement au moindre coût est une stratégie que l'on pourrait qualifier d'optimale dans la plupart des cas. Cependant, dans des environnements non contrôlés, dynamiques et changeant – ce qui est le cas des individus évoluant dans leur milieu naturel – explorer d'autres ressources est une capacité adaptative primordiale (localiser d'autres sources de nourriture ou d'eau, repérer un itinéraire de fuite, etc.). La balance entre les comportements d'exploitation et d'exploration est un intérêt occupant une place centrale dans l'étude de la prise de décision.

Différentes politiques permettent d'expliquer pourquoi les individus ne choisissent pas toujours les options les plus optimales a priori. Dans un système déterministe, un agent finit par converger vers un même comportement et réagir de la même façon aux évènements de l'environnement. L'apprentissage par renforcement classique tend donc vers l'application d'une politique déterministe par les agents : par essais / erreurs successifs, ils déterminent l'option la plus optimale, la répètent, et convergent ainsi vers un comportement d'exploitation. Cependant, dans des environnements où le résultat (*outcome*) des options suit une loi de probabilité ou varie au cours du temps, une stratégie déterministe résulte en général d'une mauvaise compréhension des utilités des actions possibles, et entraîne souvent un comportement sous-optimal à long terme. Les politiques stochastiques, quant à elles, permettent d'adapter le degré de variabilité des choix en fonction du contexte, et ce notamment en contrebalançant l'exploitation avec de l'exploration.

IV) Balance exploitation / exploration

Les humains, tout comme les autres animaux, dépensent un temps et une énergie non négligeables à des activités dont le but n'est pas une récompense matérielle, allant du jeu à la science, en passant par des nombreux autres comportements orientés vers un but pouvant être qualifiés de « curiosité » ou d'« exploration » (Berlyne, 1966; Burgers, 1966). De même, lors d'un choix entre plusieurs options, un individu ne choisit pas toujours l'option permettant de maximiser la récompense, il fait simplement ce choix plus fréquemment. On considère dans ce cas que choisir une option qui maximise la récompense consiste à exploiter (la ressource, l'environnement, la tâche), tandis que tout autre choix, paraissant au premier abord sous-optimal, consiste à explorer (Cohen et al., 2007). On postule alors que l'exploration consiste à expérimenter des alternatives pas ou peu connues et d'obtenir ainsi de l'information sur son environnement. En effet, baser ses choix comportementaux uniquement sur de l'exploitation est délétère dans des environnements incertains et dynamiques. Pour être capable de s'adapter à des changements dans son environnement, contrebalancer de l'exploitation pure par de l'exploration est donc bénéfique à long terme, ce qui fait d'elle une faculté indispensable à la survie.

1) L'exploration modulée par des paramètres externes

L'exploration ou la curiosité peuvent être générées spontanément, et considérées comme des processus homéostatiques : une force qui, au même titre que la faim ou la soif, pousse à agir et motive l'action. On parle alors de *drive*.

Certains stimuli a priori neutres déclenchent également des comportements d'approches ou autres comportements habituellement dirigés vers des récompenses (Berlyne, 1966). Cette capacité permet ainsi d'assurer un accès aux informations et stimuli environnementaux. L'incertitude, la nouveauté, ou la présence d'éléments saillants sont notamment des moteurs de l'exploration.

D'un point de vue éthologique, l'exploration est en général étudiée pendant les comportements de navigation spatiale au sein d'un environnement (naturel ou artificiel) (Drai et al., 2000; Dvorkin et al., 2008; Eilam and Golani, 1989). Chez l'animal notamment, ce comportement est facilement mis en jeu face à de la nouveauté ou bien vis-à-vis d'un stimulus

saillant. Un rongeur placé dans un nouvel environnement tel qu'un *open field* manifestera ces comportements d'exploration dans sa stratégie de navigation spatiale en parcourant une grande distance, en visitant de manière homogène l'ensemble de l'environnement, en se déplaçant de manière saccadée avec de multiples pauses et en exhibant des comportements et postures typiques tels que le *rearing* (s'élever sur les pattes arrière), *scanning* (mouvements de tête dans plusieurs directions) ou *sniffing* (reniflements accompagnés de mouvements des vibrisses) (Fonio et al., 2009; Maubourguet et al., 2008). De même, si dans un environnement familier à l'animal, un objet inconnu est introduit, ces comportements d'exploration seront focalisés sur l'objet.

2) L'exploration dans la sélection de l'action

Lors de choix entre plusieurs actions bien définies (appuis levier, *nose poke*, etc.), la stratégie de sélection de l'action peut également être plus ou moins exploratoire. En effet, les animaux choisissent les options les moins optimales (de prime abord) une fraction non négligeable du temps. Cette observation vient de Richard Herrnstein au milieu du 20ème siècle d'après des expériences de conditionnement opérant réalisées sur le pigeon (Herrnstein, 1961). Les animaux disposaient de deux leviers délivrant des quantités différentes de nourriture. Après conditionnement, ils appuyaient sur chacun des deux leviers de manière équivalente à leur ratio de renforcement. Herrnstein en tire une loi nommée la « *Matching Law* », qui stipule que, dans le cadre d'un choix binaire, la probabilité de choisir A est :

$$P_A = \frac{V_A}{(V_A + V_B)}$$

Les actions menant à de fortes récompenses sont donc la plupart du temps choisies, et celles menant à de plus faibles récompenses sont tout de même choisies à proportion de leur valeur relative. Par exemple, si le levier A délivre 1 dose de nourriture et le levier B 2 doses, alors l'animal choisit A un tiers des essais et B deux tiers des essais. Plusieurs explications théoriques ont été proposées pour rendre compte de la *Matching Law*, basées soit sur l'utilité (cf. Chapitre 1 – III) – 2)) soit sur la politique (cf. Chapitre 1 – III) – 3).

En général l'introduction d'une alternative nouvelle ou incertaine, ou bien d'un stimulus saillant provoque leur exploration, c'est-à-dire qu'elle présente transitoirement une valeur

subjective forte. Cela suggère donc que l'utilité de certaines options ne dépend pas que de la valeur de la récompense.

La réponse à la nouveauté peut être théoriquement abordée sous forme d'un bonus nommé *shaping bonus*. Le *shaping* consiste à apprendre un comportement à un animal par étape, en récompensant des étapes intermédiaires du comportement final. Par analogie, la nouveauté est considérée comme une récompense de substitution, ce qui explique pourquoi les animaux explorent les conséquences de nouveaux états et actions. Explorer ces nouveaux états permet éventuellement de découvrir une action intéressante. Afin que ces comportements restent temporaires, d'autres modèles proposent que le bonus de nouveauté soit intégré à la valeur d'une action comme une différence entre l'état nouveau et le précédent (Kakade and Dayan, 2002). A l'extrême, cela peut prendre la forme d'une « initialisation optimiste » : les valeurs associées aux options nouvelles sont toujours maximales, et après exploration de celles-ci, leur valeur converge à nouveau vers la récompense moyenne attendue (Kakade and Dayan, 2002).

De même, l'incertitude associée à un état, un stimulus ou à l'issue d'une action peut induire leur exploration. Une issue moins prédictible qu'une autre peut être associée à une plus grande valeur (Anselme et al., 2013; Gershman and Uchida, 2019). On parle de bonus d'incertitude. Explorer une issue incertaine permet d'acquérir de l'information, et éventuellement de réduire son incertitude. Par exemple, un animal qui explore un abri dont il connait mal les caractéristiques saura s'il est plus ou moins efficace que son abri habituel. Une distinction entre deux formes d'incertitudes a été proposée, une incertitude inattendue (surprise) déclenchant de l'exploration et une incertitude attendue (variabilité connue d'une récompense) pouvant être ignorée (Yu and Dayan, 2005) (cf. Chapitre 4). Cependant, même face à une forme irréductible d'incertitude attendue, l'exploration est utile car permet de maintenir à jour ses connaissances sur son environnement (Heilbronner et al., 2010).

Enfin, l'exploration en réponse à un stimulus saillant quel qu'il soit, même s'il n'a pas au préalable été associé à une récompense, est une réaction naturelle, parfois désignée sous le terme de généralisation, lorsque le stimulus saillant est considéré comme connu dans un autre contexte. Certains modèles suggèrent que ces stimuli saillants sont d'abord considérés par l'individu comme ambigus en termes de valence (positive ou négative), ce qui le place dans un état d'incertitude ne pouvant être résolu que par une collecte d'information via une augmentation transitoire d'attention (Kakade and Dayan, 2002).

Néanmoins, des comportements d'exploration sont observés même en dehors de toute notion de nouveauté ou de saillance. Plusieurs études montrent qu'il existe une exploration

résiduelle dans les choix de nombreuses espèces. Or, la plupart du temps, la balance exploration/exploitation est alors étudiée dans des paradigmes dans lesquels les récompenses sont délivrées de manière stochastique (Anselme et al., 2013). Cela suggère que l'exploration est toujours présente en tant que *drive*. Si explorer systématiquement est, tout comme exploiter systématiquement, sous-optimal, il s'agit de « trouver un compromis » entre les deux politiques. On parle de « balance » (ou compromis, *trade-off*) entre exploitation et exploration. On peut voir cela comme un problème sur deux échelles de temps : exploiter localement est plus rentable à court terme, et explorer permet de s'adapter sur des échelles de temps plus longues. Plusieurs façons de régler cette balance chez les animaux ont été proposées, pour que le niveau d'exploration soit intégré dans la politique comme un choix plus ou moins de stochastique.

Une première stratégie (« *ε-greedy* », cf. Chapitre 4 pour une définition plus formelle) consiste à encourager les comportements d'exploration de manière aléatoire, au hasard, en intégrant du « bruit » dans la décision (Wilson et al., 2014a). En d'autres termes, l'option optimale est exploitée la majeure partie du temps, et l'option alternative est explorée de temps en temps de manière aléatoire. Bien que généralement peu optimale, cette politique fonctionne néanmoins à moindre coût cognitif.

Une autre stratégie est de considérer les décisions comme fondamentalement aléatoires mais biaisées par l'utilité. C'est une généralisation de la *Matching Law*, la proportion des choix reflétant les utilités relatives. L'exploration y est dirigée par la différence relative de l'ensemble des valeurs des options disponibles. Suivant un formalisme plus fin que la simple *Matching Law*, le gain du *trade-off* exploration/exploitation (ou sensibilité à la valeur, ou degré de bruit dans la décision) est ajustable en fonction de l'environnement ou de l'état de l'individu. Cette stratégie d'exploration, appelée *Softmax* (cf. Chapitre 4 pour une définition plus formelle) permet à long terme d'assurer à l'organisme de bonnes facultés d'adaptation tout en disposant de suffisamment de ressources, et est donc très souvent utilisée pour modéliser les décisions en neurosciences. Par ailleurs, adopter cette stratégie permet de biaiser les décisions vers de la recherche d'informations, et de promouvoir dans un certain degré les options les moins certaines via des bonus. Ce type de stratégie mixe donc exploration aléatoire (politique stochastique) et exploration dirigée (bonus d'information dans l'utilité) (Wilson et al., 2014a).

V) Conclusions

Dans ce premier chapitre, j'ai présenté les trois principaux effets que produit une récompense : elle provoque une sensation de plaisir, permet d'apprendre des comportements par association, et construit des buts vers lesquels elle génère de la motivation. Ces trois fonctions sont au cœur des processus de prise de décision, ce qui fait de la récompense un composant majeur dans l'étude de ces derniers. La plupart des études cherchant à caractériser les bases neurales de la prise de décision se concentrent sur l'aspect d'apprentissage que les récompenses permettent, en résumant des choix à des valeurs d'associations apprises par expérience. Bien souvent, les paradigmes expérimentaux mis en place consistent à apprendre au sujet (humain ou animal) qu'il doit déclencher un choix au moment où un stimulus particulier est présenté. On dit qu'ils sont « guidés par des indices » (ou *cue-guided* en anglais). Cela présente l'avantage de synchroniser le comportement – et potentiellement l'activité neuronale – à un instant précis choisi par l'expérimentateur, permettant ainsi d'étudier et de modéliser les politiques de décision des sujets. Ces cadres d'études considèrent généralement les processus de prise de décision à l'état stationnaire et comme uniquement basés sur des comparaisons de valeurs de récompenses. Les aspects liés à l'hédonisme et surtout à la motivation sont davantage laissés de côté, et les comportements auto-générés sont mal compris. Comme nous l'avons vu pourtant, les comportements anticipatoires ou d'approche peuvent être déclenchés sans l'intervention d'indices externes, uniquement guidés par le but que représente une récompense. En milieu naturel – hors cadre d'expérimentation – nous admettons effectivement que certaines décisions sont orientées par des stimuli environnementaux discrets et saillants, mais aussi en grande majorité sont orientées de manière spontanée sans déclencheurs particuliers.

Plusieurs questions demeurent peu abordées. Comment un animal initie de lui-même un comportement dirigé ? Comment les comportements auto-générés peuvent être étudiés dans le cadre de la prise de décision ? Par quels processus neuronaux une récompense permet d'orienter un comportement auto-généré et une stratégie de choix ? Ce sont ces questions qui ont guidé le travail effectué durant ma thèse. Les prochains chapitres seront consacrés à la présentation des circuits cérébraux étudiés, avant d'exposer les résultats de ces travaux.

Chapitre 2 : Le circuit méso-cortico-limbique

Le circuit méso-cortico-limbique a une fonction prépondérante dans l'encodage des récompenses et les processus de prise de décision. Dans ce chapitre, je présenterai le système dopaminergique (DA) en me focalisant sur l'Aire Tegmentale Ventrale (VTA). Je décrirai en particulier comment les propriétés électrophysiologiques des neurones DAergiques soutiennent l'encodage des récompenses dans les processus de choix. Les rôles des cortex orbitofrontal (OFC) et préfrontal (PFC) seront exposés dans le chapitre suivant.

I) Le système dopaminergique

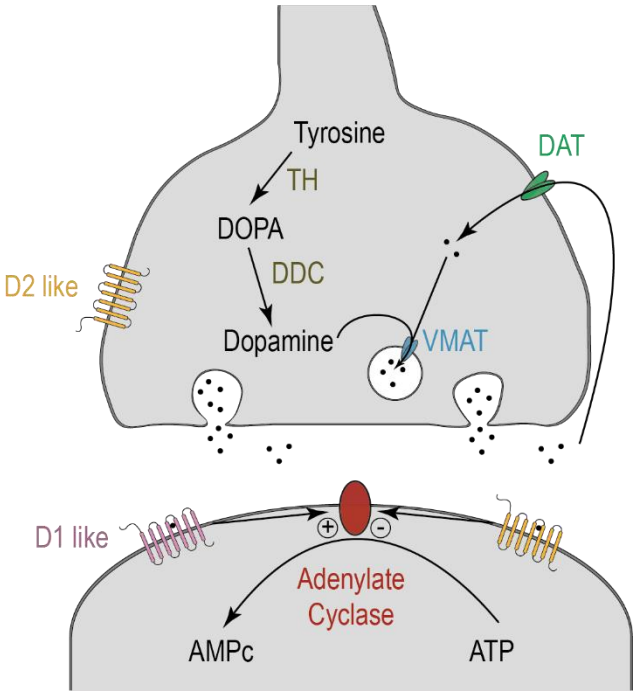
1) Dopamine et neurotransmission

Dans les années 1920, le physiologiste et pharmacologue Otto Loewi établit que les neurones communiquent entre eux par l'intermédiaire d'un mécanisme chimique que l'on désigne aujourd'hui sous le terme de neurotransmission. L'existence de diverses formes de neurotransmission monoaminergique fut découverte par la suite en 1964 par Dahlström et Fuxe (1965), qui identifièrent des neurones contenant de l'adrénaline, noradrénaline et dopamine.

Synthétisée à partir de la Tyrosine sous l'action de la Tyrosine Hydroxylase (TH), la dopamine (DA) appartient à la famille des catécholamines, tout comme l'adrénaline et la noradrénaline issues de la même voie de biosynthèse (Meiser et al., 2013) (Figure 4). La présence de l'enzyme TH dans les neurones permet donc d'y détecter la présence de catécholamines. La DA est stockée dans des vésicules sous l'action du transporteur vésiculaire des monoamines VMAT, et est transportée dans ces vésicules jusqu'à la membrane du neurone lors de la libération. La libération de DA n'a pas forcément lieu au niveau d'une synapse, puisque des terminaisons axonales de neurones DAergiques sans complexes protéiques synaptiques ont été identifiées (Descarries et al., 2008), on parle alors de « varicosités » et de conduction volumique (Liu et al., 2021). Contrairement aux neurotransmetteurs qui véhiculent un signal sur une échelle spatio-temporelle restreinte, la fonction des neuromodulateurs réside dans une diffusion plus large au niveau de populations de neurones, ainsi que dans une régulation plus lente liée aux propriétés des récepteurs auxquels ils se lient (Cools, 2019). Bien

que l'action de la DA fut longtemps considérée comme homogène au sein d'une structure cible, elle semble pouvoir cibler des zones relativement restreintes et fluctuer selon une constante de temps de l'ordre de la centaine de millisecondes (Patriarchi et al., 2018).

L'action de la DA est médiée par l'activation des récepteurs dopaminergiques qui appartiennent à la famille des récepteurs transmembranaires couplés aux protéines G, récepteurs métabotropiques induisant des signalisations intracellulaires lentes. Deux types de récepteurs DAergiques sont distingués d'après leur première identification : les *D1-like* et les *D2-like*, qui respectivement augmentent ou diminuent l'activité de l'adénylate cyclase (AC) (Girault and Greengard, 2004; Kebabian and Greengard, 1971; Liu et al., 2021; Tritsch and Sabatini, 2012) (Figure 4). La famille des récepteurs D1 regroupe les sous-types D1 et D5, couplés à une protéine Gs activatrice de l'AC, et la famille des récepteurs D2 les sous-types D2, D3 et D4 couplés à une protéine Gi/o inhibitrice de l'AC. Sous l'action de l'AC, l'adénosine triphosphate (ATP) intracellulaire est convertie en adénosine monophosphate cyclique (AMPc) dont l'augmentation en concentration active la protéine kinase A (PKA). L'activation de cette dernière entraîne, par phosphorylation, plusieurs cascades de réactions intracellulaires avec différents effecteurs selon les structures et types cellulaires (Tritsch and Sabatini, 2012), induisant généralement une augmentation de l'excitabilité intrinsèque et synaptique du neurone.



<u>FIGURE 4 :</u> Synthèse et signalisation dopaminergique
La DA est synthétisée par le neurone dopaminergique à partir de la tyrosine via l'action de la tyrosine hydroxylase (TH) et de la dopamine décarboxylase (DDC). Elle est accumulée au sein de vésicules par le transporteur vésiculaire des monoamines (VMAT). Une fois libérée dans l'espace extracellulaire, elle agit sur les récepteurs métabotropiques *D1-like* ou *D2-like* ayant des actions antagonistes sur l'adénylate cyclase (AC). L'AC catalyse la synthèse du second messager AMPc qui initie une cascade de signalisation intracellulaire. La DA peut aussi être recapturée du milieu extracellulaire par le transporteur de la DA (DAT).

Au contraire, une diminution de la concentration d'AMPc réduit l'activité de la PKA ce qui favorise, parmi d'autres effecteurs, l'ouverture des canaux K+ activés par les protéines G (canaux GIRK), entraînant ainsi une réduction de l'excitabilité neuronale (Beaulieu & Gainetdinov 2011). Les récepteurs DAergiques sont exprimés de manière distribuée dans le cerveau, avec une prépondérance dans le circuit mésocorticolimbique (voir plus bas), et peuvent être localisés tant au niveau somato-dendritique que terminal des neurones (Liu et al., 2021). Les récepteurs DAergiques de type D2 sont également souvent exprimés par les neurones DAergiques eux-mêmes (on parle alors d'autorécepteurs), permettant de réguler négativement l'activité et donc le relargage de DA (Sesack and Picke, 1994).

La DA en excès dans le milieu extra-cellulaire peut être recapturée par les neurones grâce au transporteur de la DA DAT, ou bien être dégradée par la catécholamine-O-méthyl-transférase (COMT) (Bilder et al., 2004). Ces deux mécanismes participent à restreindre spatialement la diffusion de la DA, leur absence ou altération conduisant à une hyperdopaminergie.

2) Anatomie des voies dopaminergiques de l'Aire Tegmentale Ventrale

L'innervation dopaminergique dans le cerveau est essentiellement d'origine mésencéphalique. Trois noyaux mésencéphaliques sont responsables de la libération de DA : il s'agit du champ rétro-rubral (RRF), de la substance noire *pars compacta* (SNc) ainsi que de l'aire tegmentale ventrale (VTA) – noyaux numérotés respectivement A8, A9 et A10 d'après la classification anatomique de Dahlström et Fuxe (Andén et al., 1964).

Des fonctions distinctes sont classiquement attribuées à ces noyaux. Le RRF a été notamment associé aux comportements de peur (Moaddab and McDannald, 2021). Les neurones de la SNc, projetant majoritairement sur le striatum dorsal (Björklund and Dunnett, 2007; Yetnikoff et al., 2014), seraient impliqués dans le contrôle du mouvement volontaire et la formation d'habitudes (Bromberg-Martin et al., 2010). La VTA, quant à elle, est associée principalement aux fonctions liées aux récompenses et à la motivation (Bromberg-Martin et al., 2010). Cette dichotomie reste très schématique car des neurones DAergiques de la SNc répondent aux récompenses (Coddington and Dudman, 2018) et ceux de la VTA sont également impliqués dans des processus moteurs (Fischbach-Weiss et al., 2018).

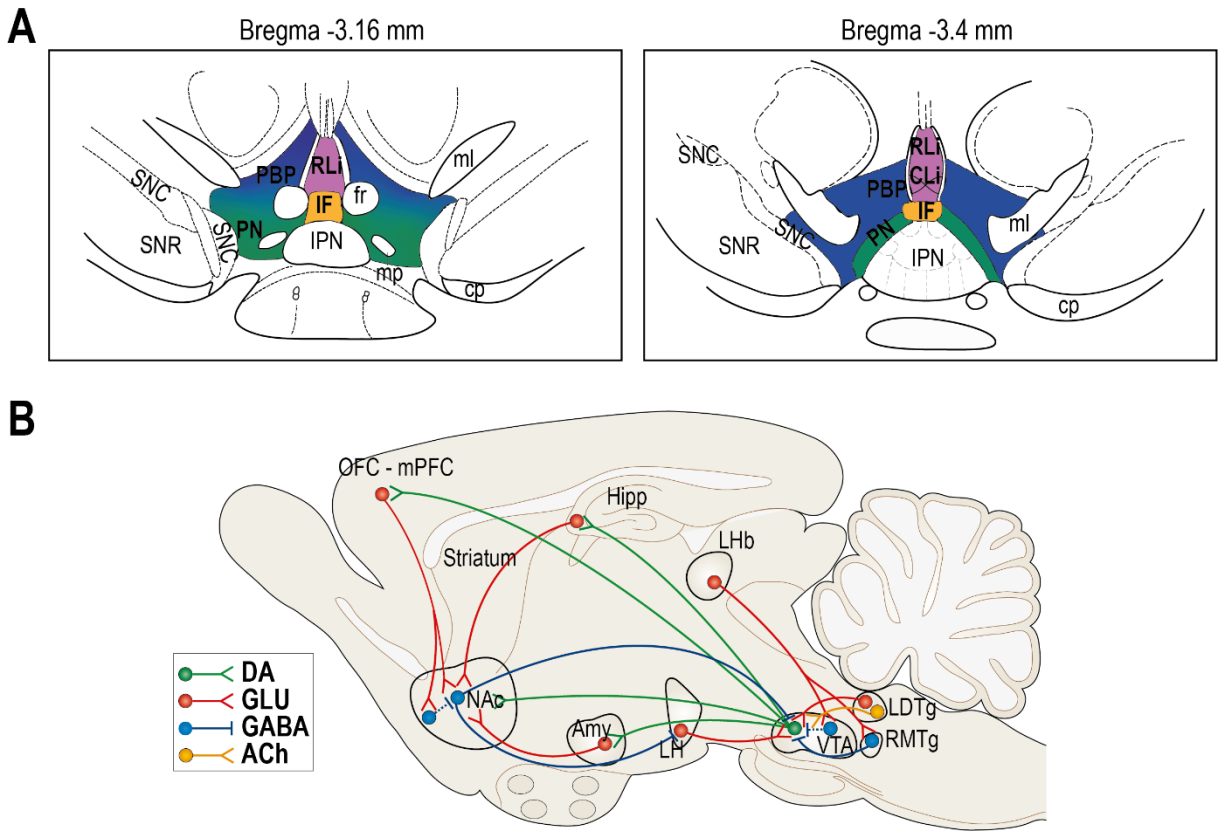


FIGURE 5 : Neuroanatomie de la VTA et du circuit méso-cortico-limbique

A. Schéma de coupes coronales de VTA de souris et ses différents noyaux à -3,1 mm (gauche) et -3,4 mm (droite) du Bregma. cp : pédoncule cérébelleux supérieur, fr : fornix, IF : noyau interfasciculaire, IPN : noyau interpédonculaire, ml : lemnisque médian, mp : pédoncule mamillaire, PBP : noyau parabrachial pigmenté, PN : noyau latéral parabrachial, RLi : noyau linéaire rostral, SNC : substance noire *pars compacta*, SNR : substance noire *pars reticulata*. *Adapté de l'Atlas Paxinos & Franklin « The mouse Brain 3rd Edition »*.

B. Représentation schématique de la connectivité entre les principales structures du circuit méso-cortico-limbique via les projections dopaminergiques (DA), glutamatergiques (GLU), GABAergiques (GABA) et cholinergiques (ACh). Amy : amygdale, Hipp : hippocampe, LDTg : tegmentum latéro-dorsal, LH : hypothalamus latéral, LHb : habénula latérale, OFC : cortex orbitofrontal, mPFC : cortex préfrontal médial, NAc : noyau Accumbens, RMTg : tegmentum rostromédial, VTA : aire tegmentale ventrale. *Adapté de Russo & Nestler 2013.*

Au cours de ce travail de thèse, je me concentrerai sur le système dopaminergique de la VTA car les circuits neuronaux qui lui sont associés sont déterminants dans les processus de motivation et de prise de décision.

Les neurones de la VTA projettent vers de multiples aires cérébrales du système limbique comme le striatum ventral (noyau Accumbens, NAc), l'amygdale (Amg) ou l'hippocampe (HPC) (Beier et al., 2015; Pistillo et al., 2015; Walsh and Han, 2014), ainsi que vers des aires corticales notamment le cortex préfrontal, orbitofrontal et cingulaire (Pistillo et al., 2015; Walsh and Han, 2014). Ces deux voies mésolimbiques et mésocorticales forment le circuit mésocorticolimbique, un vaste circuit plus communément appelé « circuit de la récompense » impliqué dans le renforcement et la gestion des récompenses. Les dérégulations du fonctionnement de ce circuit sont impliquées dans plusieurs pathologies telles que la schizophrénie, la dépression ou les addictions (Beier et al., 2015; Lammel et al., 2008; Pistillo et al., 2015).

Noyaux de la VTA et types cellulaires :

La VTA se décompose en plusieurs noyaux et sous-régions (Figure 5A) : le noyau parabrachial pigmenté (PBP) qui correspond à la plus large partie de la VTA, le noyau paranigral (PN) situé plutôt latéralement, et les noyaux interfasciculaire (IF), linéaire rostral (RLi) et linéaire caudal (CLi) qui sont eux situés médialement. Le noyau tegmental rostro-médian (RMTg), plus récemment appelé « queue de la VTA » en raison de sa forme, est considéré comme faisant partie de la VTA (Ikemoto, 2007).

Des études basées sur la révélation immunohistochimique de la TH ou des hybridations in situ d'ARNm (Hnasko et al., 2012; Margolis et al., 2006; Nair-Roberts et al., 2008; Olson and Nestler, 2007; Swanson, 1982), et plus récemment l'utilisation d'outils optogénétiques (Morales and Margolis, 2017) ont permis d'évaluer la répartition anatomique et les types neuronaux composant la VTA. La VTA est donc composée de trois types neuronaux différents. Les neurones DAergiques, situés en majeure partie dans le PBP et le PN, représenteraient environ 60% des neurones de la VTA. Environ 30% des neurones de la VTA seraient des neurones GABAergiques (libérant du GABA, acide γ-amino-butyrique, neurotransmetteur inhibiteur). Ces neurones peuvent être soit des neurones de projection, soit des interneurones inhibiteurs, dont une partie est issue notamment du RMTg et confèrent une inhibition interne à la VTA (Jhou et al., 2009). Enfin, moins de 10% des neurones de la VTA seraient des neurones glutamatergiques (libérant du glutamate, GLU, neurotransmetteur excitateur), et seraient situés majoritairement dans les noyaux médiaux (Hnasko et al., 2012).

Par ailleurs, certains neurones peuvent co-relarguer de la DA et d'autres neurotransmetteurs. En effet, un co-relargage de DA et GLU a été mis en évidence au niveau du NAc d'abord par observations en microscopie électronique, puis par des études

immunohistochimiques révélant la colocalisation du transporteur vésiculaire du GLU (VGLUT) et de la TH (Descarries et al., 2008). Plus récemment, des enregistrements électrophysiologiques en tranches de neurones GABAergiques du NAc couplés à des activations optogénétiques de neurones DAergiques par la *channelrhodopsine - 2* (ChR2) l'ont confirmé (Stuber et al., 2010; Zhang et al., 2015). Ce co-relargage serait issu non seulement d'un même neurone mais plus précisément de la même fibre axonale (Zhang et al., 2015). L'hypothèse d'un co-relargage depuis une même vésicule est en revanche débattue (Zhang et al., 2015). Au niveau du PFC, des études anatomiques par hybridation in situ et marques immunohistochimiques indiquent qu'il existe également une telle cotransmission (Yamaguchi et al., 2011). L'activation optogénétique de neurones DAergiques par la ChR2 couplée à des enregistrements électrophysiologiques en tranches de neurones du striatum a également révélé un co-relargage de DA et de GABA (Tritsch et al., 2012). S'il a été observé chez le rat des neurones DAergiques coexprimant la TH et l'enzyme de synthèse du GABA (GAD), ce n'est pas le cas chez la souris. Par ailleurs, certains neurones GABAergiques ne relarguant pas de DA exprimeraient également la TH. Une hypothèse alternative est que certains neurones DAergiques exprimeraient un transporteur de GABA (GAT) afin de capturer le GABA depuis l'espace extra-cellulaire. Le transporteur vésiculaire des mono-amines VMAT semble par ailleurs capable de stocker le GABA dans les vésicules des neurones DAergiques (Tritsch et al., 2012). Les neurotransmetteurs GLU et GABA possèdent des cinétiques d'action temporellement plus rapides et spatialement plus restreintes que la DA car se lient à des récepteurs spatialement organisés sur l'élément post-synaptique. Les rôles fonctionnels d'une telle co-transmission à plusieurs échelles spatio-temporelles restent obscurs, mais il est proposé que cela permettrait de réguler le signal transmis de manière plus fine et flexible au sein de circuits anatomiquement fortement interconnectés.

Connectivité de la VTA :

De multiples afférences, efférences et sous-circuits au sein du système mésocorticolimbique ont été décrits à ce jour (Figure 5B), dans un premier temps par des études histologiques et anatomiques, et désormais de plus en plus par des stratégies de traçage viral couplées à des études électrophysiologiques et comportementales, qui tentent d'associer à chaque connexion neuronale un rôle fonctionnel particulier.

Le **noyau Accumbens** (NAc), situé dans le striatum ventral, est une cible majoritaire des projections DAergiques de la VTA (Lammel et al., 2012, 2008; Morales and Margolis, 2017; Pistillo et al., 2015; Swanson, 1982; Walsh and Han, 2014).

Les projections DAergiques de la VTA agiraient majoritairement sur les neurones épineux moyens (MSN) du NAc *shell* et *core* qui sont GABAergiques (Pistillo et al., 2015). Des travaux utilisant des traceurs rétrogrades (Watabe-Uchida et al., 2012), ou bien couplant outils optogénétiques et enregistrements électrophysiologiques ont mis en évidence des connexions directes GABAergiques depuis ces MSN sur les neurones GLUergiques et GABAergiques de la VTA (Xia et al., 2011). Des études basées sur des analyses de traçages antérogrades et rétrogrades proposent qu'il existe une interface de connexions réciproques entre les neurones DAergiques du mésencéphale et les différentes régions du striatum, formant une spirale organisée topologiquement le long d'un axe médio-latéral (pour les noyaux DAergiques) et ventro-dorsal (pour les structures cibles) (Beier et al., 2019; de Jong et al., 2019; Haber et al., 2000). Les neurones DAergiques de la VTA médiale projetteraient vers le NAc ventromédial, qui projetterait en retour vers les régions plus latérales de la VTA, laquelle projetterait vers des régions plus centrales du striatum, et ainsi de suite jusqu'à la SN et le striatum dorsolatéral. Les connexions inhibitrices GABAergiques revenant en retour du NAc vers la VTA participeraient à l'activation de proche en proche des neurones DAergiques. Cette architecture anatomique soutiendrait une circulation de l'information neuronale depuis l'intégration limbique jusqu'à la commande motrice. Différents sous-noyaux du NAc ont été plus récemment impliqués et ces connexions réciproques décrites de manière plus détaillée. Les neurones de la partie médiale du NAc *shell* exercerait un contrôle inhibiteur sur les neurones DAergiques de la VTA, tandis que ceux de la partie latérale du NAc *shell* sur les neurones GABAergiques principalement, induisant ainsi une désinhibition des efférences DAergiques le ciblant (Yang et al., 2018).

Une autre population neuronale du NAc reçoit des afférences de la VTA, il s'agit des interneurones cholinergiques (TANs, *tonically active neurons*), capables également de libérer du GLU (Pistillo et al., 2015). Ces derniers partagent des connexions réciproques avec les MSN. Le NAc serait également la cible de projections GLUergiques et GABAergiques de la VTA (Beier et al., 2015), ainsi que le lieu d'un relargage conjoint de DA et GLU (Descarries et al., 2008; Liu et al., 2021) comme évoqué précédemment.

La voie DAergique mésostriatale est largement impliquée dans les processus d'approche (Nicola, 2010; Salamone and Correa, 2012), d'apprentissage (Bromberg-Martin et al., 2010) et de renforcement aux drogues (Chiara and Imperato, 1988; Pascoli et al., 2015).

Les neurones DAergiques de la VTA projettent également vers l'**amygdale** (Amg) (Beier et al., 2015; Lammel et al., 2008; Swanson, 1982), plus particulièrement son noyau baso-latéral (BLA). Ces projections ont été impliquées dans la formation de mémoires aversives,

notamment par conditionnement pavlovien (Tang et al., 2020), dans le contrôle des comportements de recherche de drogue (Pistillo et al., 2015; Stamatakis et al., 2014) et plus récemment dans l'anxiété notamment induite par la nicotine (Nguyen et al., 2021). En retour, la BLA modulerait indirectement l'activité de la VTA grâce aux nombreuses projections GLUergiques vers les MSN du NAc et les neurones pyramidaux des cortex frontaux tels que le PFC médian (mPFC) (Pistillo et al., 2015; Russo and Nestler, 2013), qui sont eux-mêmes connectés à la VTA.

Une autre région cible majeure des projections DAergiques de la VTA sont les **cortex frontaux**, notamment l'OFC et le PFC (Beier et al., 2019, 2015; Lammel et al., 2012, 2008). Cette voie mésocorticale a été impliquée dans la mémoire de travail, la flexibilité comportementale et la prise de décision (Floresco, 2013; Sharpe et al., 2019). Des projections directes DAergiques ont été notamment identifiées par des études de traçage rétrogrades (Beier et al., 2019; Lammel et al., 2008). Certaines de ces projections seraient à l'origine d'un co-relargage de DA et GLU (Yamaguchi et al., 2011), suggérant un rôle d'appui de la transmission GLUergique rapide dans la signalisation d'événements saillants par la DA. Les cortex préfrontaux expriment à la fois les récepteurs D1 et D2, le premier potentiellement à des niveaux d'expression plus élevés, et serait majoritairement localisé sur les cellules pyramidales (Floresco, 2013; Seamans and Yang, 2004).

Des études d'hybridation in situ d'ARNm de transporteurs vésiculaires du GLU (VGLUT) indiquent que le PFC constitue également une entrée GLUergique importante sur la VTA (Geisler et al., 2007). L'utilisation de traceurs rétrogrades couplés à des outils optogénétiques ont permis de mettre en évidence également des afférences mPFC et du cortex antérieur (ACC) vers les neurones DAergiques de la VTA (Beier et al., 2015). Ces connexions réciproques soutiennent l'idée que les cortex préfrontaux et le système DAergique interagiraient directement durant le contrôle des comportements dirigés.

Diverses connexions existent par ailleurs entre le NAc, l'Amg et le PFC. Les MSN du NAc reçoivent aussi des afférences GLUergiques depuis l'Amg ainsi que le PFC (Pistillo et al., 2015). On parle notamment pour ce dernier d'une connexion indirecte du PFC vers la VTA via le NAc. Les TAN reçoivent également des afférences GLUergiques du PFC. Les neurones pyramidaux du PFC projettent en retour vers l'Amg, ces projections auraient un rôle dans l'extinction des comportements de mémoire de la peur (Jenni et al., 2017).

Il existe des connexions réciproques entre la VTA et l'**habénula latérale** (LHb), une région épithalamique. En effet, des études de traçage antérograde depuis la LHb indiquent que les afférences GLUergiques en provenance de ce noyau contacteraient à la fois les neurones DAergiques et GABAergiques de la VTA (Lammel et al., 2012; Omelchenko et al., 2009). Par ailleurs, il existe une connexion disynaptique entre la LHb et la VTA faisant intervenir le RMTg, noyau majoritairement GABAergique, et qui serait responsable d'une inhibition de l'activité des neurones DAergiques de la VTA en absence de récompense attendue (cf. plus bas, Chapitre 2 – II) – 2)) (Hu et al., 2020; Jhou et al., 2009; Maroteaux and Mameli, 2012). Les efférences de la VTA vers la LHb sont principalement GLUergiques (Geisler et al., 2007; Morales and Root, 2014). La LHb signalerait en particulier les stimuli aversifs, et participerait au traitement des récompenses en relation avec l'état interne (Hu et al., 2020; Lammel et al., 2012). Ses connexions avec le système DAergique seraient déterminantes dans la régulation homéostatique des comportements dirigés vers les récompenses.

La VTA reçoit par ailleurs des afférences de nombreuses structures et de divers types. Une de ses majeures entrées provient des **noyaux cholinergiques du pont**, tels que le noyau tegmental latéro-dorsal (LDTg) (Dautan et al., 2016; Lammel et al., 2012; Russo and Nestler, 2013), et dans une moindre mesure le noyau tegmental pédonculo-pontin (PPTg), qui projette néanmoins préférentiellement sur la SNc. L'innervation cholinergique depuis ces structures encapsule en quelque sorte la VTA et ses projections (Omelchenko and Sesack, 2006). Des études couplant outils optogénétiques et enregistrements électrophysiologiques chez la souris montrent que les afférences cholinergiques du PPTg et du LDTg modulent l'activité des neurones DAergiques et GABAergiques de la VTA (Dautan et al., 2016). La signalisation cholinergique provenant de ces noyaux agit notamment sur les récepteurs nicotiniques de l'ACh largement exprimés dans la VTA par les neurones DAergiques et GABAergiques. Cette transmission module à la fois l'activité de ces neurones (Naudé et al., 2015; Tolu et al., 2013) mais aussi la prise de décision sous incertitude (Naudé et al., 2016).

Certains neurones GLUergiques de la VTA projettent également vers diverses structures telles que le NAc, le PFC, et l'Amg (Morales and Root, 2014), tout comme certains neurones GABAergiques projetant (Beier et al., 2019, 2015). La VTA reçoit aussi des afférences GLUergiques du LDTg (Lammel et al., 2012), de l'hypothalamus latéral (LH) (Geisler et al., 2007; Watabe-Uchida et al., 2012), et du noyau basal de la strie terminale (BNST, *basal nucleus*

of the stria terminalis) (Glangetas et al., 2015). Elle est également soumise à une modulation sérotoninergique en provenance du raphé dorsal (Beier et al., 2015).

La richesse des connexions au sein du circuit méso-cortico-limbique soutient sa fonction d'intégration d'informations à la fois environnementales et émotionnelles ainsi que d'activation de comportements motivés et orientés.

3) Activité spontanée des neurones dopaminergiques et relargage

La connectivité fournie des neurones DAergiques de la VTA s'accompagne d'une dynamique particulière de leurs modes de décharge. L'activité spontanée des neurones DAergiques de la VTA a été caractérisée dans les années 1990, d'abord *in vivo* par des techniques d'enregistrements intra et extracellulaires chez le rat anesthésié (Grace and Bunney, 1984a, 1984b), puis chez le primate éveillé durant des tâches comportementales simples (Schultz, 1986). Cette description a été complétée par des études menées *in vitro* sur tranche par la technique de *patch-clamp* (Shi, 2009). Ces travaux décrivent une activité lente (entre 2 et 10 Hz) et irrégulière qui est le continuum entre deux rythmes distincts décrits plus bas (Figure 6A).

Trois types d'activité sont décrits :

- Un **état inactif** correspondant à un état hyperpolarisé du neurone dans lequel aucun PA n'est émis. Le neurone est alors dit « silencieux ». Cet état s'observe *in vivo* lorsque les neurones DAergiques sont sous l'influence d'afférences inhibitrices en provenance du pallidum ventral (VP) (Floresco et al., 2001). Une population de neurones DAergiques silencieux représente autant de neurones mobilisables et activables, ce qui confèrerait au système DAergique une meilleure réactivité face à des stimuli saillants dans l'environnement.

- Un **mode de décharge lent et régulier**, observable *in vivo* et également *in vitro* en l'absence des afférences, et provenant d'oscillations du potentiel de membrane du neurone (Grace and Bunney, 1984a). On parle ainsi d'activité *pace-maker* ou intrinsèque, générée par des conductances sensibles aux variations de potentiel de membrane. Le potentiel d'action (PA) active des canaux calciques voltage-dépendants (VGCC), dont l'entrée de calcium en résultant

active les canaux SK perméables au potassium. La sortie d'ions K+ aboutit à l'hyperpolarisation du neurone laquelle active les canaux HCN, responsable d'un courant de rectification entrant dit I_h, lequel repolarise progressivement le potentiel de membrane pour le ramener au seuil de déclenchement du PA. Cette repolarisation est entraînée et contrebalancée par respectivement, les VGCC et les canaux potassiques voltage-dépendants (Kv) (Shi, 2009). Cette activité *pace-maker* serait associée à un relargage tonique régulier de DA dans les structures cibles (Floresco et al., 2003) (Figure 6B), maintient une concentration de DA extracellulaire relativement stable. Ce tonus DAergique est régulé au niveau somatique par les récepteurs D2 de haute affinité qui, par auto-inhibition, exercent un contrôle négatif sur les neurones DAergiques (Shi, 2009).

- Un **mode de décharge soutenu et transitoire en bouffées de PA**. Une bouffée, aussi dit *burst*, est initiée par une succession de deux PA espacés de moins de 80 ms et se termine lorsque deux PA espacés de plus de 160 ms, le second ne participant alors pas à la bouffée (Grace and Bunney, 1984b) (Figure 6A). La fréquence des PA dans une bouffée peut alors atteindre 50 Hz. Les PA au sein d'une bouffée sont généralement d'amplitude décroissante. Cette activité n'est pas observable sur les enregistrements *ex vivo* en tranche ce qui suggère qu'elle est une propriété émergente du réseau grâce aux diverses afférences que reçoit le noyau DAergique. Les bases moléculaires du *burst* sont multiples. Il dépendrait non seulement d'entrées excitatrices GLUergiques via les récepteurs NMDA et cholinergiques en provenance des noyaux du pont (PPTg, LDTg), mais également de conductances intrinsèques aux neurones (Faure et al., 2014; Paladini and Roeper, 2014). Il semblerait notamment que les canaux SK participent à la régulation des bouffées puisque leur délétion spécifiquement dans les neurones DAergiques provoque une augmentation du nombre de PA par bouffée (Soden et al., 2013). Ce mode de décharge est associé à un relargage phasique massif et transitoire de DA dans les structures cibles (Floresco et al., 2003) (Figure 6B). Le relargage phasique de DA issu de ces bouffées interviendrait en réponse à des stimuli saillants telles que des récompenses (Schultz, 1986; Schultz et al., 1997) et participerait a fortiori dans l'apprentissage par renforcement. Plus que la décharge en bouffées au niveau d'un neurone DAergique, c'est la synchronie des *bursts* et du relargage phasique au niveau de la population neuronale qui serait déterminante.

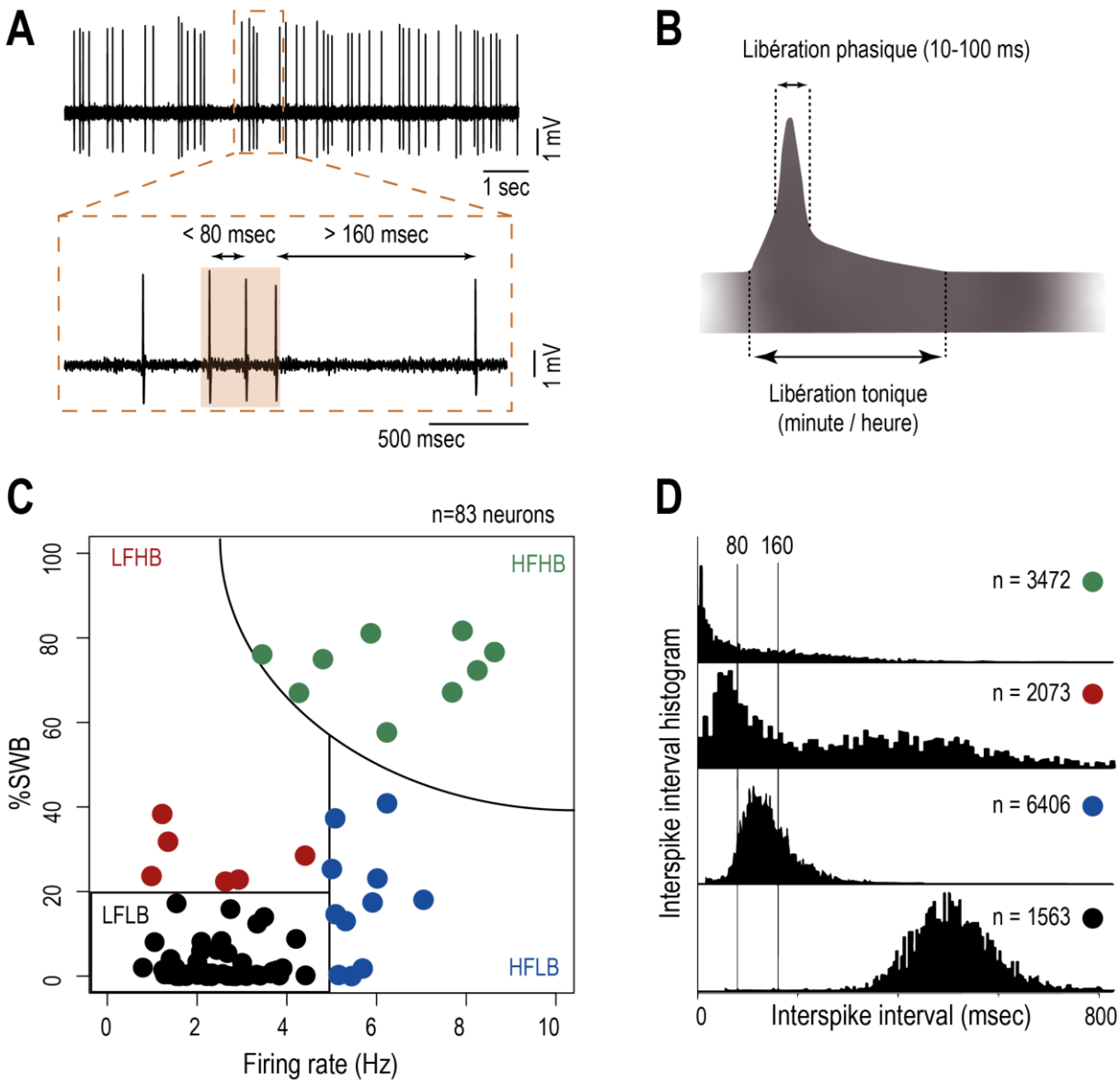


FIGURE 6 : Activité spontanée des neurones dopaminergiques de la VTA

A. Trace brute de l'activité spontanée d'un neurone DAergique enregistré en juxtacellulaire sur souris anesthésiée et décomposition de la structure d'une bouffée.

B. Schéma de la libération phasique et tonique de DA qui s'opèrent respectivement sur une courte échelle de temps (allant de 10 à 100 ms) et sur des temps plus longs (allant de plusieurs minutes à des heures). *Adapté de Schultz 2007.*

C. Diagramme des différents profils de décharge spontanée des neurones DAergiques de la VTA, en fonction de leur %SWB et de leur FR. Le profil LFLB (*low frequency and low burst firing*) correspond aux neurones ayant une faible fréquence moyenne (FR < 5 Hz) et peu d'activité en bouffées (%SWB < 20%). Le profil HFLB (*high frequency and low burst firing*) correspond aux neurones ayant une haute fréquence moyenne (FR > 5 Hz) et peu d'activité en bouffées (%SWB < 20%). Le profil LFHB (*low frequency and high burst firing*) correspond aux neurones ayant une faible fréquence moyenne (FR < 5 Hz) et une forte activité en bouffées (%SWB > 20%, en pratique entre 20% et 60%). Le profil HFHB (*high frequency and high burst firing*) correspond aux neurones ayant une haute fréquence moyenne (FR > 5 Hz) et une forte activité en bouffées (en pratique le %SWB est entre 60% et 100%). *Adapté de Faure 2014.*

D. Exemples de distributions d'intervalles entre PA (en msec) pour un neurone DAergique de chaque classe (HFHB, HFLB, HFLB, LFLB). *Adapté de Mameli-Engvall 2006.*

L'activité en bouffée est un mode de décharge qui vient s'ajouter à l'activité *pace-maker* du neurone DAergique (Figure 6A), en association à des entrées synaptiques particulières. Néanmoins tous les neurones DAergiques ne font pas la même quantité de *bursts*. Pour décrire plus finement les profils d'activité des neurones DAergiques, deux principales mesures sont utilisées : la fréquence de décharge (*firing rate, FR*) moyenne du neurone et le pourcentage de PA contenus dans une bouffée (*percentage of spikes within burst, %SWB*) (Figure 6C). Sur la base de ces mesures, différents profils de décharge spontanée peuvent être identifiés (Figure 6C-D).

L'émission en bouffées est sous la dépendance de plusieurs régulations multiples qui peuvent être affectés par des facteurs environnementaux et génétiques. En effet, les neurones DA n'émettent pas ou peu de bouffées in vitro, quand leurs innervations ont été coupées (Faure et al., 2014; Grace and Bunney, 1984a). Par exemple, la suppression de certains récepteurs normalement exprimés par les neurones DAergiques a tendance à modifier globalement l'activité de ces derniers. La délétion du récepteur nucléaire aux glucocorticoïdes (GR) spécifiquement dans les neurones DAergiques induit une diminution de leur activité, à la fois en fréquence moyenne et en %SWB (Barik et al., 2013), sur animal anesthésié. La délétion constitutionnelle de la sous-unité β2 (animaux $\beta2^{-/-}$) du récepteur nicotinique à l'acétylcholine (nAChR) aboutit sur animal anesthésié à une quasi-abolition du mode de décharge en bouffées, mais la ré-expression de la sous-unité spécifiquement dans la VTA restaure les divers profils de décharge des neurones DAergiques (Faure et al., 2014; Mameli-Engvall et al., 2006). Ces résultats, comme beaucoup d'autres, suggèrent que diverses régulations hormonales et neuromodulatrices contrôlent l'activité DAergique.

L'exposition chronique à des substances d'abus peut également modifier les profils d'activité des neurones DAergiques. Des souris soumises à une consommation passive et chronique de nicotine présentent une activité DAergique globalement plus élevée, à la fois en fréquence moyenne et en %SWB (Morel et al., 2018; Tolu et al., 2017). La consommation active d'alcool à faibles doses semble être associée à un effet similaire (Juarez et al., 2017). D'autres facteurs environnementaux modifient de manière durable l'activité des neurones DAergiques de la VTA. Des états d'anxiété et de stress social provoqués par des protocoles de défaite sociale sont associés chez la souris à une activité DAergique spontanée plus élevée, à la fois en fréquence moyenne et en %SWB (Barik et al., 2013; Morel et al., 2018). Enfin, il semble que l'activité spontanée des neurones DAergiques soit liée à l'expression de traits de personnalité individuels tels que des profils de prise de décision (Torquet et al., 2018).

Si divers facteurs semblent participer au contrôle de l'activité DAergique, celle-ci ne détermine pas systématiquement les niveaux de DA circulant dans le milieu extracellulaire (Mohebi et al., 2019). En effet ces niveaux de DA peuvent également être régulés via des mécanismes de recapture par exemple, ou bien de rétrocontrôle agissant directement sur la synthèse ou libération de DA (Liu et al., 2021). Ainsi, une hyperdopaminergie a été reliée à de l'hyperactivité et des troubles attentionnels et l'hypodopaminergie à des états catatoniques et des pathologies telles que la schizophrénie (Faure et al., 2014).

Dans l'ensemble, la neuromodulation DAergique affecte l'activité des structures cibles de différentes manières. Il est suggéré que son action dans le striatum affecterait l'intégration des entrées synaptiques corticales sur les MSN via son action sur les récepteurs D1 et D2 (Surmeier et al., 2007; Tritsch and Sabatini, 2012), l'activation des premiers favorisant un état stable plus dépolarisé, en augmentant l'excitabilité du neurone, et celle des seconds l'inverse. L'action des récepteurs D1 favoriserait à la fois la transition vers des états de forte activité (parfois appelés « up-states ») et le maintien dans cet état. Ces actions semblent participer au maintien des comportements motivés. Une absence de DA semble être responsable d'oscillations pathologiques dans les ganglions de la base (Jenkinson and Brown, 2011). Dans les cortex, l'action de la DA affecterait l'activité oscillatoire de réseaux neuronaux (Puig and Miller, 2012), en particulier en stabilisant des états de forte activité (Durstewitz et al., 2010) qui sous-tendent des rôles dans l'attention et la mémoire de travail. Les mécanismes selon lesquels la DA corticale participe aux prises de décisions restent plus obscurs. L'effet de la DA sur les cortex est décrit plus détail dans le Chapitre 3. En modifiant la transmission synaptique et les propriétés intrinsèques des neurones cibles dans le cortex et le striatum (mais aussi dans l'hippocampe, l'amygdale, etc.), la DA exercerait une action globale sur le système nerveux central (Berke, 2018).

L'ensemble des actions de la DA sur ces structures cibles impacte les comportements à différentes échelles de temps, que je décrirai dans la prochaine partie en me focalisant sur les aspects liés aux récompenses et à la prise de décision.

II) Dopamine, récompense et prise de décision

La diversité des modes de régulation de la neuromodulation DAergique, que ce soit au niveau de l'activité des neurones DAergiques, du relargage ou de la recapture, participe à l'idée que la DA joue de multiples rôles. Les premières fonctions associées à la DA concernent le contrôle du mouvement volontaire et découlent d'observations cliniques de pathologies liées à des déficits du système DAergique telles que la maladie de Parkinson. Par la suite, la DA s'est révélée comme un acteur clé dans l'apprentissage, à la fois dans les comportements conservés sur le plan évolutif, dans la prise de décision, mais également dans des pathologies comme l'addiction. Dans cette partie, je développerai le rôle de la DA dans les fonctions liées aux récompenses.

1) Le système dopaminergique et la stimulation intracrânienne

Une étude majeure a permis de faire le lien entre le système dopaminergique et les effets renforçants d'une récompense. Il s'agit des travaux de James Olds et Peter Milner, qui, dans les années 1950, explorent les aires cérébrales pouvant être impliquées dans la construction de représentations positives ou négatives de stimuli. Ils procèdent pour cela à l'implantation d'électrodes de stimulation qu'ils introduisent dans divers noyaux et aires cérébrales de rats (Olds and Milner, 1954), et placent ces derniers dans des boîtes de Skinner. Les animaux peuvent ainsi presser librement sur un levier pour s'auto-administrer des stimulations électriques de la région cérébrale implantée. Ils découvrent ainsi que plusieurs aires cérébrales présentent des propriétés suffisamment renforçantes lorsqu'elles sont stimulées pour induire l'acquisition de ce comportement opérant. Parmi ces régions majoritairement sous-corticales, se trouvent notamment l'hypothalamus, le thalamus, le tegmentum et le septum (Olds and Milner, 1954). D'autres régions sont identifiées comme ayant des effets aversifs, et d'autres encore des effets neutres. Ces régions ayant des propriétés renforçantes sont pour la plupart reliées entre elles par le faisceau médian du télencéphable (MFB) (Olds and Olds, 1963; Wise, 1996). Le MFB est un faisceau composé de nombreuses fibres afférentes et efférentes de divers types reliant le télencéphale et le mésencéphale, dont une majorité provient de la VTA et de la SNc. Des fibres cholinergiques et GLUergiques semblent également emprunter le MFB

(Shizgal, 1997; Wise, 2005). Il innerve plusieurs régions telles que l'hypothalamus, le striatum et les cortex frontaux. Le MFB est une des régions cérébrales les plus efficaces qui, stimulée électriquement, induit un renforcement. L'ensemble des structures identifiées par Olds et Milner est rapidement considéré comme un circuit plutôt que comme des régions impliquées de manière isolée, et le concept de « circuit de la récompense » est alors énoncé. Ce dernier semble être conservé dans de nombreuses espèces vertébrées, ce qui s'accompagne de la vision évolutionniste qu'un circuit sous-tendant les émotions ou expériences hédoniques liées aux récompenses ou aux punitions participe à améliorer la survie (Panksepp, 2011).

Par la suite, l'utilisation d'antagonistes de la DA dans des expériences d'auto-stimulation du MFB a conduit à considérer la DA comme candidate principale, dans ce circuit, de la représentation cérébrale de la récompense (Wise, 2005). En découle la théorie classique quant à l'effet de la stimulation électrique du MFB dans les protocoles de renforcement et d'auto-stimulation : une stimulation de ce faisceau de fibres induirait un relargage massif de DA notamment dans le NAc, et mimerait ainsi l'effet d'une récompense. La stimulation du MFB, en détournant en quelque sorte le fonctionnement naturel du système de la récompense, serait alors utilisée pour induire un apprentissage. Elle ne déclenche pas de comportement spécifique (comme pourrait le faire la stimulation d'une zone motrice), mais elle crée une représentation de récompense (et potentiellement de but), ainsi qu'une utilité associée, qui pousse l'animal à travailler pour l'obtenir (Shizgal and Conover, 1996). Cependant, ces expériences de renforcement par relargage de DA ne tranchent pas entre une représentation hédonique de la récompense, une composante motivationnelle, ou un signal d'apprentissage (Berridge, 2007).

2) L'erreur de prédiction de récompense

Les premières expériences corrélant plus directement l'activité des neurones DAergiques et l'encodage d'une récompense sont des enregistrements *in vivo* sur singe éveillé réalisés par Wolfram Schultz (Schultz, 1986; Schultz et al., 1997). Ces enregistrements avaient pour but de tester l'implication motrice des neurones DAergiques du mésencéphale (à la fois situés dans la VTA et la SNc). Pourtant, dans des comportements stimulus-réponse, les neurones DAergiques s'activent au moment du stimulus plutôt qu'au moment de la réponse motrice (Schultz, 1986).

Ces associations stimulus-réponses étant conditionnées, cela a conduit les auteurs à étudier les réponses des neurones DAergiques à la récompense, pendant les différentes étapes du conditionnement (Ljungberg et al., 1992) (Figure 7A). Les auteurs remarquent qu'avant apprentissage, les neurones DAergiques augmentent transitoirement leur fréquence de décharge en réponse à des récompenses primaires (nourriture ou boisson). Après apprentissage, ils répondent à des stimuli saillants non contrôlés de l'expérience, tels que l'ouverture d'une porte - qui est effectivement un stimulus conditionné prédisant la distribution d'une récompense. Schultz propose donc que l'activité DAergique n'est pas seulement une réponse à un stimulus appétitif (une « représentation de la récompense »), mais peut refléter également l'anticipation d'une récompense.

Cela le conduit à étudier la réponse des neurones DAergiques lors d'un conditionnement pavlovien (Schultz et al., 1997). Avant apprentissage, lorsqu'un jus de fruit (US) est donné au singe, les neurones DAergiques répondent de manière phasique par une augmentation transitoire de leur activité. Y est ensuite associé un stimulus lumineux neutre, qui précède systématiquement d'une seconde la distribution du jus de fruit. Après de nombreuses répétitions de cette association, les neurones DAergiques ne s'activent plus en réponse à la récompense, mais au stimulus conditionné (CS). Enfin, une fois cette association renforcée, le singe est soumis à une omission de récompense. Le stimulus lumineux s'allume, mais une seconde plus tard, le jus de fruit n'est cette fois pas délivré. Les neurones DAergiques s'activent toujours au moment du CS, mais leur fréquence de décharge diminue au moment où aurait dû être délivré l'US. Cette inhibition lors de l'omission serait le reflet d'une erreur par rapport aux associations que l'animal à apprises.

Cette expérience montre que les neurones DAergiques ne répondent pas uniquement à l'aspect hédonique d'une récompense. Au contraire l'activité DAergique est formalisée comme une erreur de prédiction de récompense (RPE) (Schultz et al., 1997), plus précisément par l'algorithme de « différence de trace temporelle » (Sutton and Barto, 1998). Entre les essais, la RPE est simplement la différence entre la récompense obtenue et la récompense prédite :

$$d = r - v$$

Avec *d* la RPE, *r* la récompense obtenue et *v* la valeur de la récompense prédite.

Cela permet d'expliquer pourquoi avant apprentissage, aucune association n'ayant encore été formée, les neurones DAergiques s'activent seulement à la récompense inattendue (la RPE est positive). Après conditionnement, la RPE au moment de la récompense est nulle, car la prédiction correspond à ce qui a été obtenu. Lors de l'omission, la RPE au moment de la

récompense attendue est négative car l'animal n'a pas obtenu la récompense, ce qui résulte en une diminution d'activité DAergique au moment de l'omission.

L'activation des neurones au moment du stimulus s'explique par la différence de trace temporelle (TD), un algorithme d'apprentissage par renforcement développé par Sutton et Barto dans les années 1990 (Sutton and Barto, 1998). Dans ce modèle, la RPE est la différence entre la récompense obtenue *plus la prédiction des valeurs futures* (décomptées d'un facteur γ), moins la valeur prédite :

$$d(t) = r(t) + \gamma * v(t+1) - v(t)$$

Avant apprentissage, le stimulus ne prédit pas de valeur future et le terme en $\gamma * v(t+1)$ est nul, alors qu'après apprentissage, le stimulus prédit une récompense. La ressemblance entre l'activité des neurones DAergiques et une TD-RPE a conduit à postuler que l'activité phasique DA serait plutôt un signal d'apprentissage.

3) L'erreur de prédiction de récompense comme signal d'apprentissage par renforcement

Cette ressemblance entre DA et RPE suggère que ces neurones participent à la règle de correction proposée par Rescorla et Wagner dans leur modèle de conditionnement pavlovien (cf. Chapitre 1 – I) – 2)) vingt ans plus tôt (Montague et al., 2004; Schultz, 1998), et affinée par Sutton et Barto sous la forme du *TD learning* (Sutton and Barto, 1998). La force d'une association stimulus-réponse est modulée d'un essai à un autre grâce à cette correction : si l'erreur est positive, l'association est renforcée, mais si elle est négative c'est le contraire.

Depuis, d'autres études ont également observé des corrélats de RPE dans l'activité des neurones DAergiques chez diverses espèces animales (Bayer and Glimcher, 2005; Cohen et al., 2012). Le transfert de ce pic d'activité des neurones DAergiques du moment de l'US vers le CS semble s'accompagner également du transfert d'un relargage transitoire de DA dans le NAc. En effet, des expériences menées en voltamétrie chez le rat indiquent que l'augmentation de DA extracellulaire enregistrée dans le NAc en réponse à une récompense survient au moment du CS après apprentissage (Day et al., 2007; Hart et al., 2014). Cela renforce l'hypothèse que l'activité en bouffée des neurones DAergiques jouerait dans l'apprentissage par renforcement par le biais d'un relargage massif et phasique de DA dans le NAc.

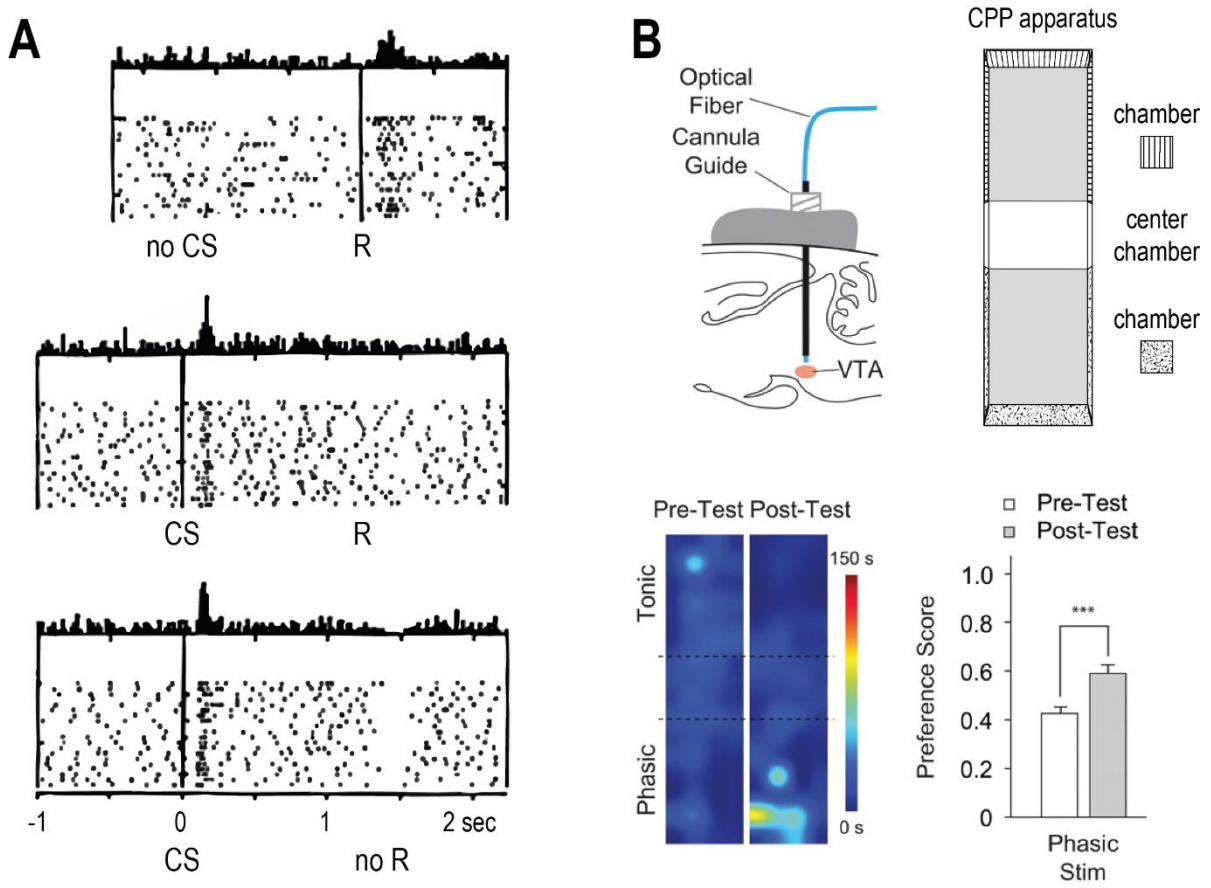


FIGURE 7 : Les neurones dopaminergiques encodent une RPE

A. Exemple de l'activité d'un neurone DAergique enregistré chez un singe à différentes étapes du conditionnement simple. Les histogrammes représentent l'activité neuronale moyennée sur plusieurs essais. Dans les graphes du dessous, chaque ligne constitue un essai et chaque point un PA. Haut : La récompense (R) est délivrée de manière inattendue. Milieu : Après apprentissage, le CS prédit la distribution de la récompense et celle-ci est délivrée. Bas : Le CS prédit la récompense mais cette dernière est omise. *Adapté de Schultz 1997.*

B. Expérience de manipulation optogénétique de l'activité des neurones DAergiques. Haut : Schéma de l'expérience. La fibre optique est implantée au niveau de la VTA chez des souris TH-Cre dont les neurones TH+ expriment la ChR2. La manipulation a lieu durant une tâche de conditionnement de préférence de place (CPP), dans une chambre de CPP dont chaque compartiment est associé à un type de stimulation, phasique (50 Hz) ou tonique (1 Hz). Bas : Temps passé dans le compartiment ayant été associé à la stimulation phasique après conditionnement. Score de préférence (pourcentage de temps passé dans la chambre associée à la stimulation phasique sur le temps total passé dans les deux chambres) avant et après le conditionnement. *Adapté de Tsai 2009.*

La manipulation causale de l'activité DAergique par des outils optogénétiques a confirmé ce rôle (Tsai et al., 2009). Dans cette étude du groupe de Karl Deisseroth (Figure 7B), la ChR2 est exprimée spécifiquement dans les neurones DAergiques de la VTA de souris TH-Cre à l'aide d'une stratégie dépendante de Cre-lox, leur permettant d'activer spécifiquement ces neurones par stimulation optique. Dans une chambre de conditionnement de préférence de place (CPP), ils associent l'un des deux compartiments avec une stimulation dite phasique à 50Hz

des neurones DAergiques de la VTA. Le jour du test, les animaux présentent une large préférence pour ce compartiment, ce qui confirme le rôle causal de l'activité en bouffée des neurones DAergiques dans l'apprentissage par renforcement. Par ailleurs, comme avec la stimulation électrique du MFB, les rats sont capables d'apprendre des comportements opérants (appuyer sur un levier par exemple) pour obtenir des stimulations optogénétiques activatrices de neurones DAergiques de leur VTA (Steinberg and Janak, 2013).

Il a aussi été montré qu'activer ces neurones par stimulation optogénétique permet de contrer l'effet de blocage (cf. Chapitre 1 – I) – 2)).

$$A \rightarrow \text{récompense}$$

$$A + B \rightarrow \text{récompense} + \text{stimulation des neurones DAergiques}$$

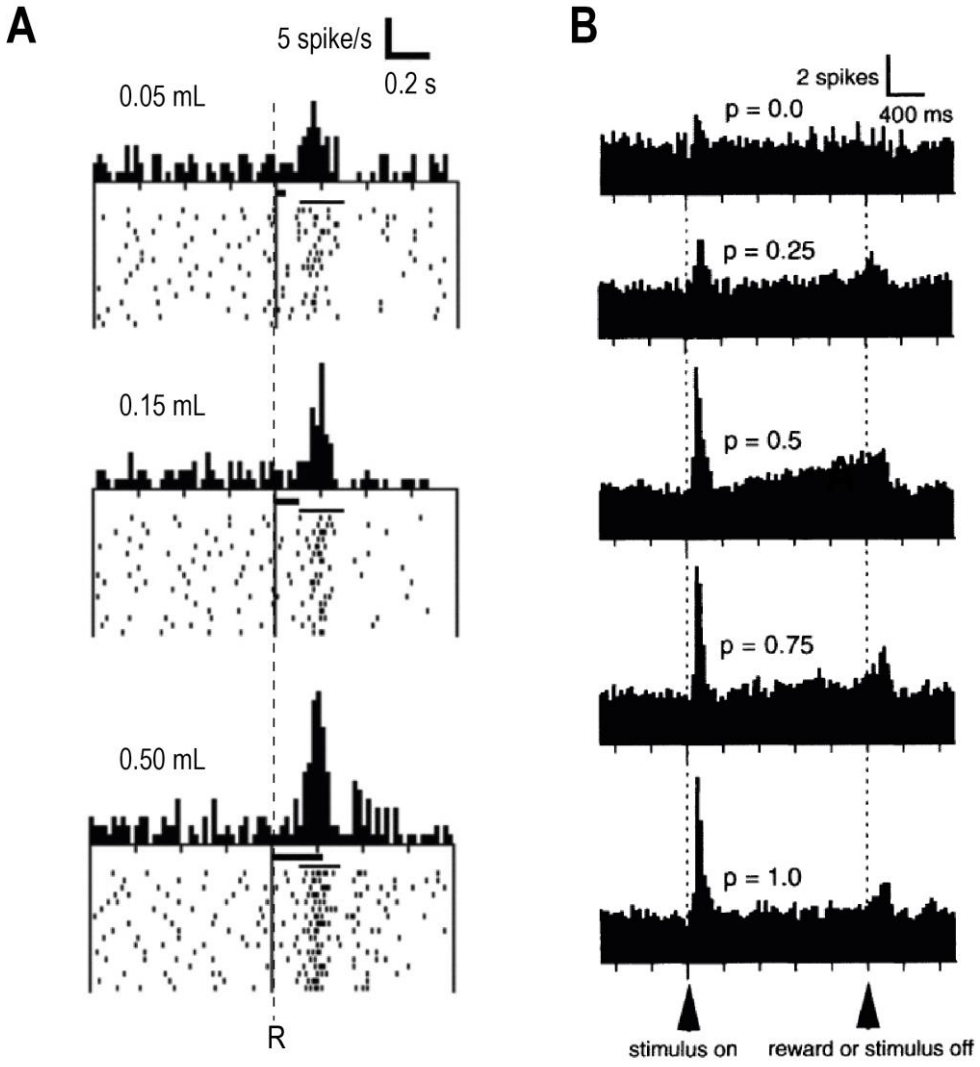
$$B \rightarrow \text{réponse}$$

L'augmentation de l'activité DAergique ainsi provoquée au moment de la récompense induit un apprentissage sur le stimulus B, apprentissage normalement bloqué par le premier CS A (Steinberg et al., 2013). Puisque le stimulus A prédit déjà la récompense, il n'y a pas de RPE au moment de la récompense et cela « bloque » l'apprentissage sur le stimulus B. Ajouter une stimulation des neurones DAergiques au moment de la récompense créerait donc une RPE positive (« la récompense obtenue est mieux que prévue »), et une association entre le stimulus B et une valeur positive aurait lieu. Cette expérience confirme que l'activité DAergique est causalement impliquée dans ce calcul de RPE et ne reflète pas simplement l'obtention d'une récompense.

4) L'erreur de prédiction de récompense comme signal quantitatif de valeur

Dans les modèles, le terme de RPE est une différence quantitative entre la quantité de récompense et la valeur prédite, qui converge vers l'espérance de la récompense (Schultz, 1998; Sutton and Barto, 1998). Pour investiguer l'encodage de la valeur d'une récompense par l'activité DAergique, il faut moduler la magnitude de ces dernières, c'est-à-dire faire varier par exemple leur quantité ou leur probabilité d'être délivrées.

Le groupe de Schultz a continué son exploration du codage de l'activité DAergique chez le singe en utilisant différentes quantité d'une même récompense (jus de fruit) (Tobler, 2005). Dans ce paradigme, des stimuli visuels (CS) distincts sont associés aux différents volumes de récompense distribués.



<u>FIGURE 8 :</u> Corrélats de valeur et d'incertitude dans l'activité dopaminergique

A. Exemple de l'activité d'un neurone DAergique enregistré chez un singe en réponse à des récompenses (R, jus de fruit) délivrées avec 3 volumes différents (0.05 mL, 0.15 mL et 0.5 mL) de manière pseudo-aléatoires et en absence de tout stimulus prédicteur. Les histogrammes représentent l'activité neuronale moyennée sur plusieurs essais. Dans les graphes du dessous (*raster plots*), chaque ligne constitue un essai et chaque point un PA. Les barres horizontales épaisses au-dessus des *raster plots* désignent respectivement le temps de la distribution de la récompense. *Adapté de Tobler 2005.*

B. Activité de la population de neurones DAergiques enregistrés chez un singe après un conditionnement simple dans lequel le stimulus prédit une récompense (jus de fruit) avec différentes probabilités d'être délivrées (p = 0, 0.25, 0.5, 0.75 et 1). Les histogrammes sont une moyenne de tous les essais (récompense délivrée et omise) de tous les neurones enregistrés. *Adapté de Fiorillo 2003.*

Ils observent que le pic prédictif d'activité DAergique au moment du CS reflète la quantité de récompense associée. Par ailleurs, avant association avec un CS, la réponse des neurones DAergiques est plus importante pour des plus gros volumes de récompense, indiquant une discrimination neuronale de la quantité de récompense (Figure 8A). De même, dans le cadre

d'un conditionnement pavlovien probabiliste, ce sont les probabilités de récompense *p* de survenir qui varient (*p* variant de 0.25 à 1) (Fiorillo et al., 2003). Après conditionnement, l'activité DAergique phasique reflète effectivement ces probabilités (Figure 8B). L'activité prédictive au moment des CS croit avec la probabilité *p* : plus le CS annonce une récompense de forte probabilité, plus l'activité prédictive est forte. L'inverse est observé au moment de la récompense : plus la récompense est attendue avec une forte probabilité *p*, plus la réponse DAergique au moment de son obtention est faible. Des travaux menés chez la souris rapportent les mêmes corrélations entre amplitude de l'activité prédictive DAergique et la quantité (Cohen et al., 2012) ou probabilité (Eshel et al., 2016) de récompense.

D'autres travaux ont mis en évidence que le délai temporel entre un CS et l'obtention de l'US influence l'activité prédictive et la réponse des neurones DAergiques (Kobayashi and Schultz, 2008; Roesch et al., 2007). Après association, plus le délai prédit par le CS est grand, plus l'activité DAergique prédictive est faible et plus la réponse au moment de la récompense est grande.

5) Incertitude liée à la récompense

L'étude de l'activité DAergique dans un cadre probabiliste a également permis d'étudier la représentation de l'incertitude, qui consiste alors en la variance de la probabilité *p* de récompense. Autrement dit, une récompense ayant une probabilité nulle ou certaine de survenir (*p = 0* ou *p = 1*) sera associée à une incertitude nulle, tandis qu'une probabilité *p = 0,5* sera associée à une incertitude maximale. Dans l'étude de Fiorillo *et al.* (Fiorillo et al., 2003) (Figure 8B), les auteurs observent également à l'échelle de la population, une augmentation d'activité DAergique de manière soutenue et durable reflétant l'incertitude associée à la distribution de la récompense. Cette « rampe » d'activité durant les quelques secondes précédent l'US corrèle en intensité avec l'incertitude qui y est associée, est donc maximale pour une probabilité *p* de 0,5. Des mesures de DA dans le NAc par voltamétrie cyclique chez le rat indiquent également qu'après apprentissage, un relargage soutenu et durable de DA pendant la présentation d'un CS corrèle en intensité avec l'incertitude associée à la récompense délivrée (Hart et al., 2015). La libération soutenue et prolongée de DA dans le striatum a également été reliée à la réduction de la distance séparant l'animal d'une récompense dans un labyrinthe (Howe et al., 2013). Si l'existence de cette rampe d'activité a fait foi durant quelques années, elle est néanmoins remise

en question car ces observations n'ont pas toujours été répliquées. Certains auteurs l'observent chez le singe particulièrement lorsqu'un CS annonce graduellement la valeur de la récompense, et suggèrent un rôle de signal motivationnel (Wang et al., 2021).

L'origine de cette rampe d'activité soutenue des neurones DAergiques est mal comprise et débattue, car ne s'inscrit pas bien dans le cadre des algorithmes d'apprentissage par renforcement du *TD learning*.

6) L'erreur de prédiction de récompense comme monnaie commune

D'autres études menées chez la souris ne rapportent que l'incertitude soit encodée sous la forme d'une rampe, mais proposent que celle-ci est intégrée dans l'activité phasique des neurones DAergiques de la VTA (Dabney et al., 2020; Fiorillo, 2011; Lak et al., 2014), qui encoderait sous la forme d'une monnaie commune à la fois la valeur et l'incertitude d'une récompense. L'idée que le signal phasique de RPE intègre différents paramètres explique que les neurones DAergiques signalent à la fois une erreur de prédiction de récompense et qu'ils sont activés par des stimuli nouveaux, incertains ou saillants qui n'ont pas encore été associés à des récompenses (Kakade and Dayan, 2002). Certains auteurs proposent d'ailleurs que les neurones DAergiques encodent une erreur de prédiction « généralisée » qui signalerait, pour tout aspect important pour le comportement, la divergence entre le résultat réel et attendu (Gardner et al., 2018).

Par extension, l'encodage de la RPE par les neurones DAergiques pourrait constituer une représentation de valeur subjective, plutôt que de dimensions objectives de la récompense (Lak et al., 2014). Des études indiquent que cette activité de RPE des neurones DAergiques n'est pas corrélée de manière linéaire à la magnitude des récompenses choisies, ce qui suggère que plusieurs paramètres sont encodés dans cette activité neuronale (Stauffer et al., 2014). Cela suggère une implication dans les décisions basées sur la comparaison de valeur (cf. Chapitre 1). Ces représentations sont donc étudiées dans des cadres actifs de prise de décision, notamment dans des tâches où l'animal fait face à un choix binaire.

Chez le singe, l'un des paradigmes très utilisés est une tâche dans laquelle l'animal indique son choix en fixant son regard sur l'un des deux indices visuels présentés à lui sur un écran. Chaque indice mène classiquement à des récompenses associées à différentes probabilités ou quantités, ou éventuellement à des punitions. Les préférences de l'animal

permettent de déduire la valeur subjective qu'il attribue à chaque stimulus, et ainsi d'étudier l'activité neuronale dans un cadre de choix économique. Des expériences similaires existent chez le rongeur dans lesquelles sont proposés à l'animal plusieurs leviers sur lesquels appuyer ou ports dans lesquels mettre son museau (*nosepoke*) afin d'obtenir différentes récompenses, le début d'un essai étant indiqué par un indice sonore, visuel ou olfactif (*cue*). Dans ce type de tâches guidées par un indice sensoriel, des études concluent que l'activité des neurones DAergiques est modulée à la fois par le type de récompense, leur quantité, le risque associé (Fiorillo, 2011; Lak et al., 2014; Stauffer et al., 2014) mais aussi le caractère informatif d'un CS (Bromberg-Martin and Hikosaka, 2009). L'activité phasique prédictive des neurones DAergiques au moment du choix est d'autant plus importante que c'est l'option préférée (donc avec la valeur subjective la plus grande) qui est sélectionnée par l'animal. Cela soutient l'idée que l'activité phasique DAergique intègre différentes dimensions de récompenses en une valeur reflétant l'utilité de la récompense, on parle de « monnaie commune » permettant de comparer sur une même échelle d'utilité des dimensions physiquement incommensurables.

Plusieurs études menées chez la souris et le singe ont également mis en évidence que l'activité des neurones DAergiques peut signaler des stimuli aversifs. S'ils encodent des stimuli saillants appétitifs sur le plan comportemental, cela semble également être le cas des stimuli portant des valences motivationnelles négatives. En général, ces derniers consistent en de brusques souffles d'air dirigés vers la tête de l'animal (*air puff*), ou bien de courtes décharges électriques administrées aux pieds de l'animal par le biais d'un plancher électrifié (*foot shock*). Certains neurones DAergiques de la VTA répondent à ces punitions, par une augmentation phasique de leur activité (Cohen et al., 2012; Eshel et al., 2016), d'autres par une réduction de leur activité (Matsumoto and Hikosaka, 2009). Suite au conditionnement de ces stimuli aversifs à des stimuli neutres, certains neurones présentent même une activité prédictive au moment du CS (Cohen et al., 2012; Matsumoto and Hikosaka, 2009).

7) Neuromodulation dopaminergique et prise de décision

La neuromodulation DAergique ne peut se résumer à une transmission phasique transitoire de dizaines de millisecondes. Le tonus DAergique est également primordial pour de nombreux changements comportementaux notamment liés à des mécanismes de régulation sur des échelles de temps plus longues, comme la faim, la satiété, l'excitation, la fatigue, le stress, l'anxiété etc. En agissant sur une échelle spatiotemporelle large, la neuromodulation DAergique permet d'adapter la résultante d'un système neuronal fortement connecté, et permet ainsi à l'organisme de réagir en fonction de son état interne ou de l'environnement (Cools, 2019; Schultz, 2007). Celle-ci dépend par ailleurs de l'état basal du système, notamment via le niveau d'activité préalable des récepteurs dans les régions cibles. Des études pharmacologiques observent que lorsque le niveau d'activité basal des récepteurs est bas, leur activation par une drogue provoquera des effets « positifs », tandis que si leur niveau d'activité basal est déjà élevé, cela provoquera des effets délétères sur la performance des animaux. On parle dans ce cas de courbe dose-réponse en « U inversé ». Cet effet a été observé chez l'Homme et chez le rongeur dans des tâches de mémoire de travail, de discrimination ou de flexibilité comportementale (Cools, 2019; Floresco, 2013; Onge et al., 2011).

Plusieurs études indiquent que la modulation tonique de DA régule la motivation ou les représentations dans les procédés de décision. En particulier, une déplétion en DA dans le NAc induite par infusion d'antagoniste des récepteurs DAergiques provoque une réduction de l'effort que les animaux sont enclins à fournir pour obtenir des récompenses (Addicott et al., 2017; Salamone et al., 2009). La manipulation de la signalisation DAergique dans le PFC par infusion d'agonistes ou antagonistes des récepteurs D1 ou D2 modifie la sensibilité des animaux au risque de manière dose-dépendante et non linéaire (Floresco, 2013). Il a également été proposé que le tonus DAergique régule la dépense d'énergie selon l'état de l'environnement (Addicott et al., 2017; Beeler, 2012; Walton and Bouret, 2019). Les environnements riches favoriseraient l'exploration car, les coûts moteurs associés aux potentielles pertes (coûts d'opportunité) étant réduits, les organismes sont plus enclins à dépenser de l'énergie pour explorer. Un fort tonus DAergique pousserait donc plutôt à une dépense d'énergie, et un tonus bas à une économie d'énergie. Ainsi, les animaux ayant une délétion constitutive du transporteur DAT (DAT KO) et présentant une hyperdopaminergie explorent davantage et sont moins sensibles au coût (Addicott et al., 2017; Beeler, 2012; Costa et al., 2014). En d'autres termes, l'activité tonique DAergique, qui dépendrait notamment de l'état basal de l'individu,

exerce un contrôle « en ligne » sur la manière dont les effets à court terme de la DA sont intégrés.

Cependant, les théories sur la DA tonique proviennent de manipulations générales du système, et n'excluent pas une participation des activités phasiques. On peut ainsi se demander si les activités DAergiques phasiques de type RPE ne sont que le reflet des décisions intégrant plusieurs aspects de la récompense, ou si la DA participe à ces décisions. En effet, la théorie RPE fait un parallèle entre activité DA et signal d'apprentissage, suggérant une influence indirecte des neurones DA sur le comportement, uniquement via l'apprentissage de la valeur. D'autres auteurs (Berke, 2018; Berridge, 2012; Salamone and Correa, 2012) ont proposé que le signal DA phasique, même en étant une RPE, pouvait influencer les actions ou même les décisions en ligne, via l'action neuromodulatrice de la DA. Par ailleurs, des travaux indiquent que la RPE des neurones DA de la VTA semble intégrer une composante reliée à l'initiation volontaire du mouvement dirigé vers une récompense (Barter et al., 2015; Wassum et al., 2012). L'inhibition spécifique par optogénétique de ces neurones altère le démarrage des comportements opérants (Fischbach-Weiss et al., 2018) voire spontanés (da Silva et al., 2018), et réduit ainsi la performance des animaux. Cependant, cette vision reste débattue et d'autres auteurs indiquent que des intensités de stimulation des neurones DAergiques qui mettent en place un renforcement, échouent à initier des comportements consommatoires ou anticipatoire fins (*licking*) (Coddington and Dudman, 2018).

La question du lien entre la DA et l'initiation volontaire d'un comportement dirigé vers une récompense peut être remise au centre de l'étude de la prise de décision. En effet, la plupart des études sur la prise de décision sont menées dans des cadres où le comportement est déclenché par un indice (*cue-guided*), ce qui rend d'autant plus confuse l'interprétation quant au rôle de l'activité phasique des neurones DA au moment de cet indice : est-elle le reflet d'une décision déjà prise ailleurs dans le cerveau, ou un calcul de valeur participant à la décision à venir, synchronisé par l'événement extérieur ? L'observation et la manipulation de l'activité des neurones DAergiques de la VTA dans un cadre opérant permettrait de répondre à cette question. A travers son rôle dans l'apprentissage, la DA renforce des actions récompensées dans le passé et favorise donc l'exploitation. Néanmoins, si elle encode une monnaie commune (et donc l'incertitude, la saillance et la nouveauté), cela suggère un rôle plus vaste dans les processus d'exploration, qui reste à caractériser.

Chapitre 3 : Les cortex préfrontaux

Grâce à leur connectivité permettant l'intégration d'informations à la fois sensorielles et limbiques, et à leur organisation architecturale soutenant des dynamiques complexes, les cortex frontaux contrôlent les comportements orientés vers un but et la flexibilité comportementale. Ils participent à de nombreuses fonctions cognitives telles que la consolidation mnésique, la planification, les processus attentionnels, la prise de décision et l'exploration (Euston et al., 2012; Miller and Wallis, 2013; Rushworth et al., 2011; Wise, 2008). Dans cette partie, je décrirai l'organisation structurale, l'activité et les rôles du cortex orbitofrontal (OFC) et préfrontal (PFC) dans la prise de décision.

I) Organisation des cortex préfrontaux

1) Architecture des cortex préfrontaux

Le cortex préfrontal (PFC) et orbitofrontal (OFC) sont situés dans la partie antérieure du lobe frontal du néocortex, en avant des aires motrices et prémotrices (Carlén, 2017; Wise, 2008). Le néocortex est structuré en 6 couches parallèles (Figure 9A), contrairement à l'allocortex (cortex pyriforme ou hippocampe par exemple) qui, lui, en possède trois. Chaque couche est composée de populations neuronales différant par leur type, forme, et densité, ainsi que par leurs afférences et efférences principales (entrées / sorties). Au sein de chaque aire corticale, il existe des connexions radiales verticales entre les différentes couches, et des connexions horizontales entre les cellules d'une même couche. Les neurones possédant des fonctions similaires sont en général organisés de telle sorte qu'ils sont alignés radialement d'une couche à une autre, formant ce que l'on nomme un « colonne corticale ». Dans les aires corticales somato-sensorielles, les neurones de la même colonne corticale partagent le même champ récepteur. Dans chaque aire corticale, les afférences se font via les neurones pyramidaux qui constituent la population neuronale la plus importante du néocortex. Ces neurones GLUergiques possèdent un arbre dendritique fourni avec de longues dendrites apicales et leurs axones peuvent projeter vers des structures extracorticales, des propriétés morphologiques qui ont été associées à leurs rôles d'intégration et de transmission d'informations corticales. Ces

propriétés architecturales soutiennent une organisation fonctionnelle qui, dans le cas des cortex préfrontaux, reposerait sur une sorte de spécialisation de chaque sous-région corticale dans l'intégration d'informations plus ou moins sensorielles, motrices ou émotionnelles (Figure 9B).

En dépit de cette organisation canonique, des différences histologiques et cytoarchitecturales permettent de différencier les structures corticales. C'est d'ailleurs sur la base de ces différences que le néocortex du cerveau humain fut subdivisé en plusieurs aires par Brodmann au début du 20ème siècle (Carlén, 2017) (Figure 9C). Ces subdivisions ont été appuyées par des études fonctionnelles de lésions menées chez l'animal, notamment le singe, mais également chez le rongeur. Néanmoins, les régions préfrontales et orbitofrontales étant particulièrement développées chez l'Homme et le primate, la question de leur homologie avec celles d'autres mammifères comme les rongeurs est débattue. Au-delà de la différence macroscopique qui concerne l'absence de circonvolutions du néocortex chez le rongeur, les aires frontales chez ce dernier sont dépourvues de couche IV (on parle de cortex « agranulaire »), elles sont majoritairement « granulaires » chez le primate (ou « dysgranulaires » si la couche IV est irrégulière et fine) (Carlén, 2017; Rudebeck and Izquierdo, 2021; Wise, 2008). Cette observation remit longtemps en question l'existence de PFC et OFC chez le rongeur. Néanmoins, les régions frontales agranulaires, partagées par tous les mammifères, semblent établir des connexions neuronales avec des structures sous-corticales analogues d'une espèce à une autre. Elles projettent notamment vers le striatum ventral, et sont la cible de projections thalamiques venant du noyau médiodorsal, ainsi que d'une forte innervation DAergique. La question serait donc plutôt de savoir comment comparer les fonctions des différentes aires frontales d'une espèce à une autre (Rudebeck and Izquierdo, 2021; Wallis, 2012; Wise, 2008). Chez le primate et l'Homme, on subdivise le PFC en plusieurs régions dorsales, ventrales, médianes ou latérales, que l'on distingue par ailleurs de l'OFC situé plus ventralement et antérieurement (Figure 9C), l'ensemble étant également désigné sous le nom de cortex préfrontal. Les régions dorsales granulaires chez le primate correspondant à des représentations mentales plus complexes auraient « émergé » avec l'évolution et n'auraient, selon certains auteurs, pas forcément d'équivalent chez les autres mammifères (Carlén, 2017; Rudebeck and Izquierdo, 2021; Wise, 2008). Le lobe frontal étant moins développé chez le rongeur, on distingue principalement l'OFC, le PFC médian (mPFC), et le cortex cingulaire situé médialement et dorsalement, les aires dorsolatérales étant plutôt des aires motrices secondaires (Figure 9D).

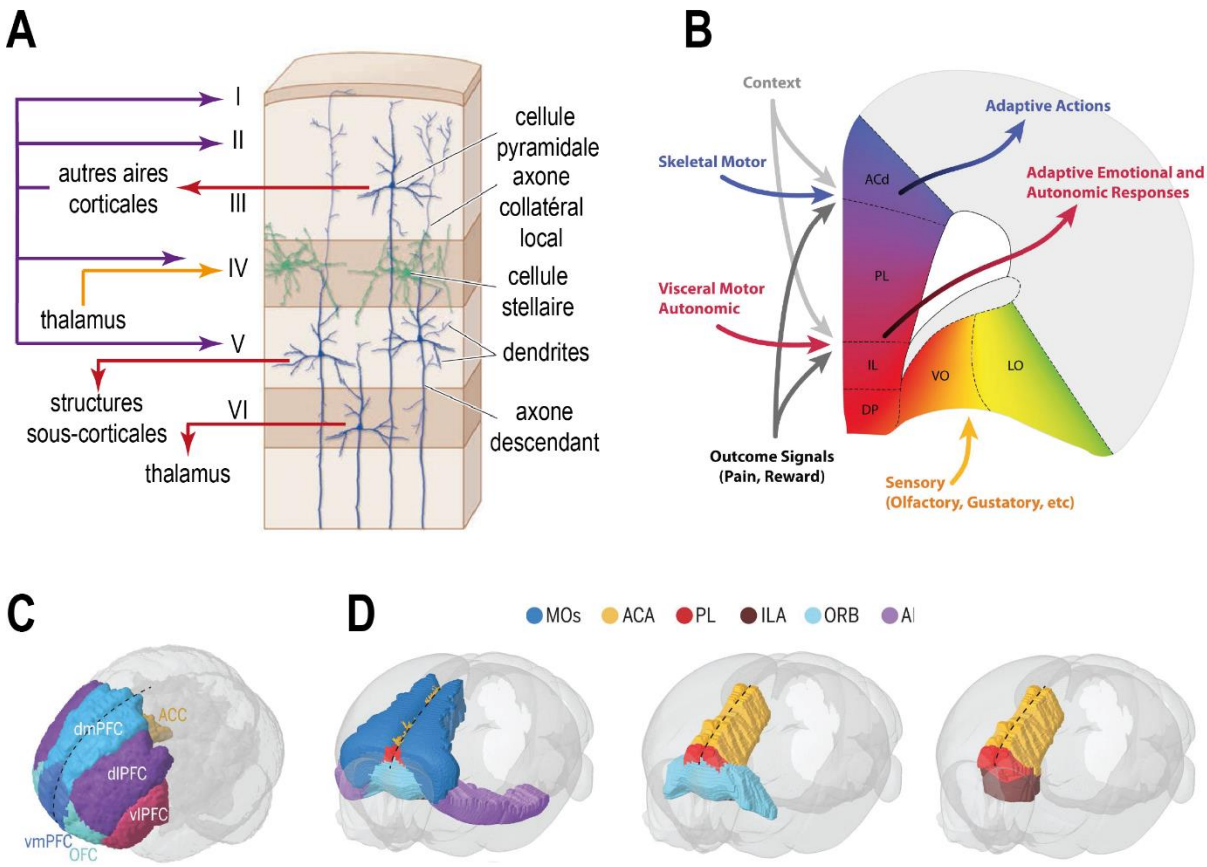


FIGURE 9 : Anatomie et organisation des cortex préfrontaux

A. Schéma de l'organisation laminaire du néocortex : La couche I (*moléculaire*) est la couche la plus externe et est en général composée de cellules gliales, ainsi que de dendrites et longs axones orientés parallèlement à la surface venant des neurones des couches plus profondes. La couche II (*granulaire externe*) et III (*pyramidale externe*) contiennent des neurones granulaires et pyramidaux ayant des connexions cortico-corticales et cortico-striatales. La couche IV (*granulaire interne*) est riche en neurones stellaires, dont les axones sont localement très ramifiés. Ces neurones reçoivent de nombreuses afférences thalamiques dans les cortex sensoriels primaires. Les couches V (*pyramidale interne*) et VI (*polymorphe*) contiennent des neurones pyramidaux plus gros dont les dendrites apicales atteignent les couches les plus superficielles et dont les axones contactent principalement des régions en dehors du cortex, notamment sous-corticales. Chaque couche contient des interneurones inhibiteurs GABAergiques qui diffèrent notamment via leurs sites d'action (somatique, dendritique, axonal). *Adapté de Purves Neuroscience 6th Edition.*

B. Représentation schématique des entrées et sorties des différentes régions du cortex préfrontal et orbitofrontal du rongeur. Cortex ACd : antérieur cingulaire dorsal, PL : prélimbique, IL : infralimbique, DP : pédonculaire dorsal, VO : orbitofrontal ventral, LO : orbitofrontal latéral. *Adapté de Euston 2012.*

C. Localisation des différentes aires du cortex préfrontal humain. dmPFC : cortex préfrontal dorso-médian, vmPFC : cortex préfrontal ventro-médian, vlPFC : cortex préfrontal ventrolatéral, OFC : cortex orbitofrontal, ACC : cortex cingulaire antérieur. *Adapté de Carlén 2017.*

D. Localisation des aires du lobe frontal chez la souris. MOs : aire motrice secondaire, ACA : aire cingulaire antérieure, PL : aire prélimbique, ILA : aire infralimbique, ORB : aire orbitofrontale, AI : aire insulaire agranulaire. *Adapté de Carlén 2017.*

Par ailleurs, il existe chez le rongeur un gradient fonctionnel dorso-ventral dans le mPFC : la partie plus ventrale, dite infralimbique (IL), semble plutôt impliquée dans le contrôle émotionnel et mnésique des comportements, tandis que la partie prélimbique (PL), plus dorsale, davantage dans la composante motrice des actions (Euston et al., 2012). Dans le cadre de la prise de décision, l'IL PFC semble principalement lié aux aspects punitifs, et le PL PFC non pas aux aspects moteurs en soi mais aux stratégies de choix (Sharpe et al., 2019). Les découpages associant chaque aire corticale à une fonction comportementale étant de plus en plus remis en question, on peut se questionner sur le type de calculs ou transformations qui sont implémentées dans ces régions préfrontales, et leurs similitudes entre espèces.

Le cortex antérieur cingulaire (ACC) est également décrit comme étant impliqué dans les processus de décision. Il signalerait notamment des erreurs de prédictions (Rushworth and Behrens, 2008), l'incertitude (Rudebeck and Izquierdo, 2021) et aurait un rôle dans la planification et l'adaptation des stratégies décisionnelles (Quilodran et al., 2008; Tervo et al., 2014). Sa participation semble néanmoins se situer en amont de la prise de décision, sur une échelle de temps longue, tandis que l'OFC et le PFC sont davantage liés à l'action et au choix en cours.

Dans ce travail de thèse, nous nous sommes donc concentrés sur l'OFC et le PL PFC en raison de leurs implications dans les comportements orientés vers une récompense, en particulier dans des contextes d'incertitude.

2) Activité oscillatoire des cortex

Le codage des récompenses et de différents paramètres orientant les stratégies de décision semble largement distribué, notamment dans le circuit mésocortical (Cisek, 2012; Hunt and Hayden, 2017; Rangel et al., 2008). Ce constat pousse à s'intéresser à l'activité de larges populations neuronales, ce qui est rendu possible par l'étude, entre autres, des variations du potentiel de champ extracellulaire (*extracellular field potential*, EFP[6]), propriété émergente du

[6] Dans la littérature, le terme « potentiel de champ local » (LFP) est très souvent utilisé, et ce pour indiquer que l'activité enregistrée correspond à une population de neurones localisés proches de l'électrode d'enregistrement. Néanmoins, puisque les fluctuations oscillatoires du potentiel extracellulaire peuvent également résulter d'une synchronisation à grande échelle d'aires cérébrales distantes, nous avons choisi d'employer le terme EFP.

néocortex permettant d'avoir accès à son activité d'ensemble. Des rythmes oscillatoires dans les cortex préfrontaux ont été impliqués dans le contrôle cognitif, rôle notamment souligné par leur altération dans des pathologies telles que la schizophrénie (Buzsáki et al., 2013; Gonzalez-Burgos and Lewis, 2008). En outre, l'activité oscillatoire a été reliée à la mémoire de travail et aux prises de décision (Amarante and Laubach, 2021; Miller and Wallis, 2013). Pour ces raisons, nous nous sommes intéressés à ces signaux oscillatoires dans l'OFC et le PFC chez la souris. Dans cette section, j'exposerai les caractéristiques principales des oscillations corticales.

Toute membrane excitable, qu'il s'agisse d'un corps cellulaire, d'une épine dendritique, ou d'une terminaison axonale, perméable aux ions et capable de faire passer des courants transmembranaires, participe à l'EFP. Ce dernier est donc formé par la sommation de nombreux événements neuronaux influençant les conductances neuronales. Les événements lents ont donc tendance à plus participer que les événements rapides qui ont tendance à s'annuler. Ainsi, les changements de potentiels dus à des entrées synaptiques, tels que les potentiels post-synaptiques excitateurs ou inhibiteurs (EPSP ou IPSP respectivement), contribuent davantage que ceux dus aux potentiels d'actions qui sont d'une durée plus transitoire.

En raison de leur morphologie particulière, l'activité synaptique des neurones pyramidaux participe fortement au potentiel de champ extracellulaire. La longue dendrite apicale favorise des entrées synaptiques sur des épines qui sont en majorité éloignées du corps cellulaire du neurone et de son axone, faisant de lui un dipôle (Figure 10A). En effet, lors d'une entrée synaptique excitatrice par exemple, l'action du GLU sur les récepteurs AMPA et NMDA provoque des courants excitateurs entrants de Na^+ ou Ca^{2+} au niveau post-synaptique. Cet influx entrant donne lieu localement à un défaut de charges positives au niveau extracellulaire, que l'on nomme *puit* (Buzsáki et al., 2012). Afin de respecter l'électroneutralité des compartiments, un efflux local du compartiment intracellulaire vers le compartiment extracellulaire se crée plus loin dans le neurone, souvent au niveau du corps cellulaire. Cette fuite de charge qui contrebalance le *puit* est nommée *source*. Ce phénomène crée donc un dipôle électrique. L'architecture corticale, par son organisation en colonne, participe à l'amplification du potentiel de champ extracellulaire. L'alignement vertical des dipôles que forment les cellules pyramidales parallèles permet la sommation des changements de potentiels dus aux entrées synaptiques, car les EPSP sont orientés dans la même direction d'une cellule à une autre. Des cellules orientées aléatoirement formeraient des dipôles orientés dans des directions opposées qui s'annuleraient potentiellement, ce qui réduirait les variations d'EFP et compliquerait l'interprétation physiologique de celles-ci (voir ci-après).

Ce sont donc en majorité les EPSP dendritiques et les IPSP somatiques (Buzsáki et al., 2012) qui contribuent à l'EFP. Il faut néanmoins souligner que des événements non synaptiques y participent également, même si dans une moindre mesure. C'est notamment le cas des états dits « bas » (*down states*) dus aux hyperpolarisations suivant l'émission d'un PA (*afterhyperpolarisation* AHP). Ces états hyperpolarisés peuvent durer plusieurs dizaines de millisecondes et leur sommation sur la population neuronale n'est pas négligeable. C'est donc la synchronie temporelle des états neuronaux et des entrées synaptiques qui modulent l'intensité de l'EFP. Si l'activité synaptique et neuronale de la région corticale enregistrée est synchrone, les fluctuations de l'EFP induites par les dipôles vont se sommer, et des oscillations de grande amplitude seront mesurées (Figure 10B). En revanche, dans les structures cérébrales où les cellules ne sont pas alignées, l'interprétation des EFP en termes de synchronie est plus ambigüe.

Ces oscillations proviennent donc de l'interaction entre des propriétés neuronales intrinsèques, les entrées synaptiques sur les cellules pyramidales, et l'architecture récurrente des connexions neuronales du cortex (Buzsaki, 2004; Buzsáki et al., 2013). Les neurones pyramidaux et interneurones GABAergiques sont fortement interconnectés au sein d'une colonne corticale. Différents mécanismes d'inhibition et désinhibition dus à ces connexions récurrentes permettent d'ajuster rapidement l'état des synapses et les dynamiques d'activité (Gonzalez-Burgos and Lewis, 2008; Hunt and Hayden, 2017; Isaacson and Scanziani, 2011). Plusieurs mécanismes interviennent dans la formation et le maintien d'activités oscillatoires. Les oscillations se formeraient sous l'impulsion d'un générateur régulier (par exemple le thalamus, la VTA, etc.) qui influencerait l'état du réseau neuronal, puis les connections réciproques du réseau de neurones corticaux permettraient l'alternance d'excitation et inhibition neuronale et donc de faire émerger et perdurer un rythme oscillatoire (Buzsaki, 2004; Buzsáki et al., 2013) (Figure 10C). Dans certains cas, les oscillations corticales peuvent également être entretenues par un « dialogue » avec une autre structure, comme c'est le cas par exemple lors des ondes lentes du sommeil qui sont maintenues par l'implication de noyaux thalamiques.

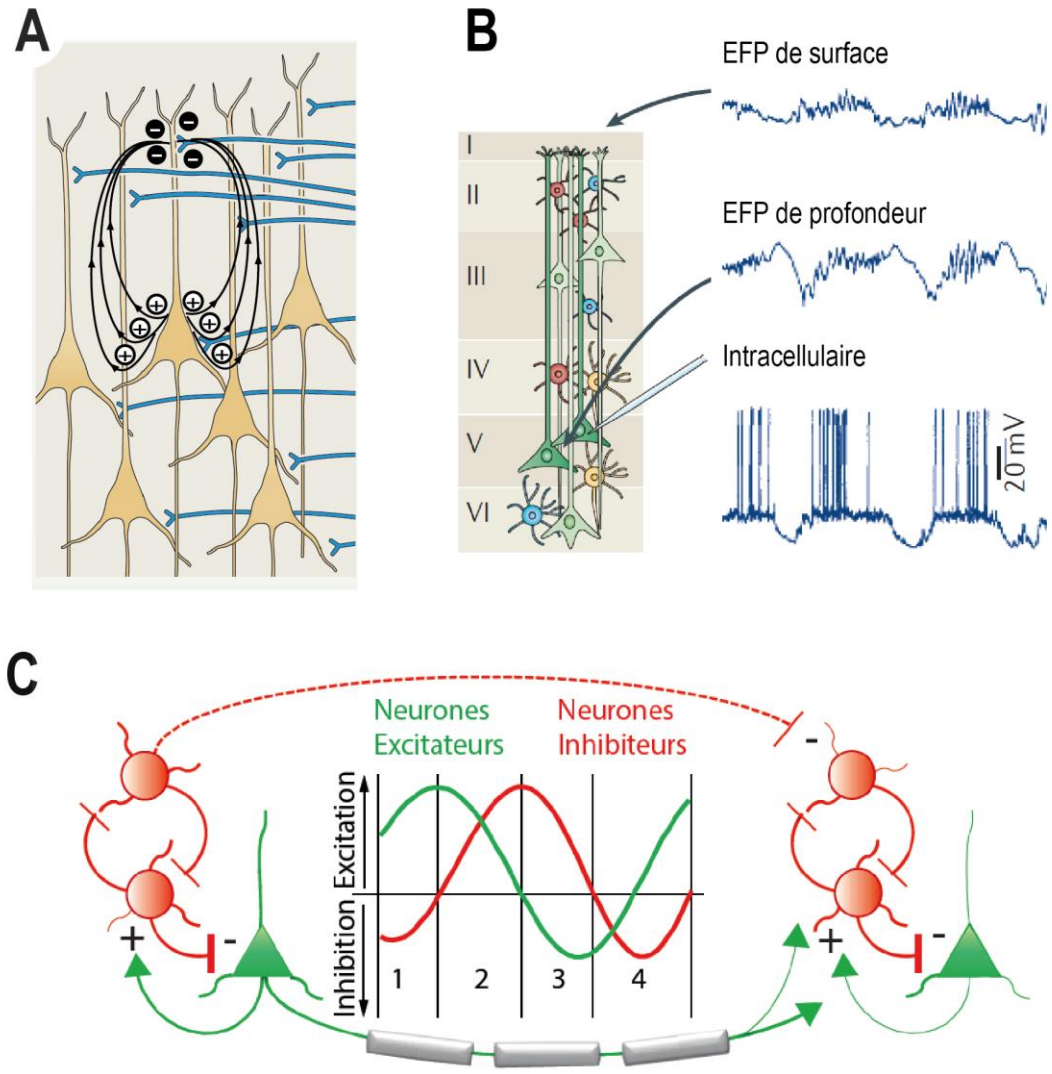


FIGURE 10 : Dipôles et oscillations

A. Représentation schématique des dipôles électriques formés par l'alignement organisé des cellules pyramidales du néocortex. *Adapté de Purves Neuroscience 6th Edition.*

B. Schéma de différentes méthodes d'enregistrement du potentiel de champ extracellulaire et intracellulaire à différentes échelles. Les traces correspondent à des enregistrements simultanés dans le cortex moteur sur chat anesthésié. L'enregistrement électrophysiologique en intracellulaire du neurone pyramidal situé en couche V indique que les alternances entre état hyperpolarisé et dépolarisé de la cellule correspondent aux variations du potentiel de champ extracellulaire enregistré en profondeur ou en surface. *Adapté Buzsaki 2012.*

C. Représentation schématique des interactions neuronales soutenant l'origine des oscillations. Ces dernières émergent de connexions récurrentes entre neurones excitateurs et inhibiteurs : (1) les neurones excitateurs cessent d'être inhibés ou sont excités, (2) ils excitent les neurones inhibiteurs, (3) les neurones inhibiteurs inhibent en retour les neurones excitateurs qui sont alors moins actifs, et (4) les neurones inhibiteurs cessent alors d'être excités par les neurones excitateurs. Et le cycle recommence. *Adapté de Freeman 1991.*

Un signal d'EFP est donc composé d'oscillations de différentes fréquences et de durées plus ou moins transitoires, qui se superposent avec différentes constantes de temps. Différentes gammes de fréquence sont distinguées. Dans un premier temps enregistrées chez l'Homme, ces gammes de rythmes oscillatoires semblent être globalement conservées à travers les espèces (Buzsáki et al., 2013), ce qui suggère leur pertinence pour l'étude des mécanismes conservés de prise de décision. Il faut néanmoins souligner que la délimitation de ces bandes de fréquence n'est pas stricte et varie légèrement d'une espèce à une autre, selon les contextes d'études et les auteurs.

Les oscillations dans la gamme de fréquence θ (7-14 Hz), très présentes dans les cortex frontaux, reflèteraient des interactions dynamiques entre aires cérébrales distantes (Cavanagh and Frank, 2014; Helfrich and Knight, 2016; Puig et al., 2014; Voytek et al., 2015). Ce rythme dans les cortex préfrontaux est à différencier du « θ locomoteur » généré par l'HPC et corrélé chez le rongeur à la locomotion ou à l'apprentissage actif (Buzsáki, 2002). Chez l'Homme, le rythme θ a été défini dans la gamme 4-8 Hz, mais semble varier chez le rongeur entre 6 et 14 Hz. Dans le PFC chez l'humain et le primate, les oscillations θ sont observées dans de nombreuses tâches cognitives par exemple la reconnaissance d'objets (Raghavachari et al., 2006), la détection de nouveauté, les tâches de mémoire de travail ou mobilisant d'autres types d'effort cognitif (Cavanagh and Frank, 2014) et ne semblent pas liées au rythme θ hippocampique. Le rythme θ frontal serait donc mis en jeu dans le « contrôle cognitif » des comportements dirigés vers un but. Chez le rongeur éveillé, le rythme θ cortical apparait principalement pendant des mouvements volontaires ou exploratoires (Buzsáki, 2002). Les ondes δ (0.5-4 Hz), associées à des états de sommeil profond, sont aussi décrites, à l'état éveillé, comme signant une synchronie distale de la VTA, le PFC et l'HPC et reflétant la sollicitation de la mémoire de travail dans un cadre de navigation orientée vers une récompense (Fujisawa and Buzsáki, 2011). Le rythme γ (30-120 Hz), souvent couplé au θ (Helfrich and Knight, 2016), émerge d'une synchronisation locale qui dépend notamment de l'interaction entre cellules pyramidales et cellules en panier (*basket cells*) exprimant la parvalbumine (PV) (Buzsáki et al., 2013). Le rythme α (8-12 Hz), est lui, associé à des états d'éveil calme chez l'humain, mais son lien avec le rythme θ, avec lequel il se recouvre, reste peu clair. Enfin, le rythme β (15-30 Hz) est représentatif d'états de haute concentration ou de feedback sensoriel (Buzsáki et al., 2013).

On considère en général que la période d'une oscillation dépend de la taille de la population neuronale engagée. Ainsi, des oscillations à haute fréquence (de courte période, dites aussi « rapides ») seraient plutôt du fait de petites populations tandis que les oscillations à

basses fréquence (de période longue, dites aussi « lentes ») recruteraient de larges réseaux neuronaux (Buzsaki, 2004). De façon complémentaire, les oscillations lentes sont reliées à des dynamiques plus stables que les oscillations rapides. Les premières correspondraient donc à une période durant laquelle une large fraction de neurones est engagée, de façon relativement synchrone, et dans une dynamique de faible dimension (état stationnaire ou cyclique). Les secondes, tout comme l'absence d'oscillations, correspondraient à des états plus désynchronisés. Ces derniers n'impliquent pas l'absence de computations mais, au contraire, des computations plus distribuées parmi les neurones, et une dynamique de plus forte dimension (plus chaotique). L'observation d'oscillations à l'échelle macroscopique reflèterait alors l'aboutissement de ces processus de calcul.

Il existe cependant des couplages entre des rythmes oscillatoires de basse et haute fréquence, c'est notamment le cas du couplage θ-γ (Helfrich and Knight, 2016). La phase de l'oscillation de basse fréquence a tendance à moduler la puissance de l'oscillation de haute fréquence à laquelle elle est couplée. Ce fonctionnement permet au cortex de traiter et de « multiplexer » différentes informations simultanément à des échelles spatiales et temporelles multiples (Helfrich and Knight, 2016).

Les oscillations corticales seraient le reflet de la formation d'assemblées neuronales et permettraient donc la sélection d'entrées synaptiques (Buzsaki, 2004). En effet, lors de variations oscillatoires de l'EFP, les conductances membranaires du neurone impliqué en seraient affectées, et la modification du potentiel de membrane de ce neurone produira un effet filtrant qui sélectionnera les entrées synaptiques sur la base de leur fréquence (Buzsaki, 2004). Une vision propose que les oscillations d'EFP participent donc à créer des assemblées neuronales flexibles. En effet, le phénomène de résonance des fluctuations de potentiel de membrane de la population neuronale modifierait la probabilité d'ouverture des canaux voltage dépendants et participerait à adapter l'excitabilité et l'activité de ces neurones (Buzsaki, 2004). Cette proposition reste néanmoins débattue car son mécanisme circulaire pose des questions de causalité entre activité neuronale et synaptique et oscillations. Dans ce travail de thèse, nous analyserons simplement les oscillations de l'EFP en tant que corrélats de l'état des assemblées neuronales, permettant d'observer l'activité d'ensemble de la structure à laquelle on s'intéresse.

II) Rôles du Cortex Orbitofrontal et Préfrontal dans la prise de décision

L'OFC et le PFC intègrent des informations à la fois limbiques et sensorielles (Euston et al., 2012; Miller and Wallis, 2013). Ils reçoivent notamment des entrées synaptiques venant de l'HPC, et sont réciproquement connectés à de nombreuses aires sensorielles corticales, relais thalamiques et régions striatales. Ils sont également soumis à une modulation neuromodulatrice notamment de la DA. Cela fait d'eux des candidats adaptés au contrôle des processus de décisions et notamment de comparaison de valeur (Padoa-Schioppa, 2011; Rushworth et al., 2011; Sharpe et al., 2019; Wilson et al., 2014b).

L'étude des structures préfrontales dans le cadre de prise de décision s'est faite, dans un premier temps, chez le primate et l'Homme. Leurs rôles comportementaux ont été déterminés notamment via des lésions, et précisés par des expériences d'enregistrements électrophysiologiques dans des tâches comportementales.

3) Le Cortex Orbitofrontal

La lésion ou l'inactivation de l'OFC induisent des déficits comportementaux qui ne sont pas simples à interpréter (Sharpe et al., 2019; Stalnaker et al., 2015). On l'a traditionnellement impliqué dans l'inhibition de l'action, et ce notamment dans des tâches d'inhibition de réponse de type *Go/No go* (Miller and Wallis, 2013; Stalnaker et al., 2015). Mais ce rôle est débattu depuis quelques années, en particulier parce que ces déficits provoqués par les lésions de l'OFC peuvent être contrés par un apprentissage plus poussé les animaux lésés (Sharpe et al., 2019; Stalnaker et al., 2015). Par ailleurs, l'enregistrement unitaire de neurones de l'OFC chez le rat dans une tâche *Go/No go* indique que leur activité n'est pas modulée par l'inhibition du mouvement (lors d'essais *No go*) mais principalement par des paramètres comportementaux (comme la direction ou le temps de réaction) (Bryden and Roesch, 2015).

La lésion de l'OFC produit également des déficits comportementaux dans des tâches d'apprentissage par inversion de contingence (*reversal learning*) (Rudebeck and Izquierdo, 2021; Rushworth et al., 2011; Sharpe et al., 2019; Stalnaker et al., 2015). Ce type d'apprentissage consiste à associer deux contextes différents (indices externes ou récompenses différentes par exemple) à des comportements différents et d'inverser les contingences après

apprentissage. Cela permet de tester notamment la mise à jour des représentations des valeurs. Un animal est capable après quelques erreurs, d'adapter son comportement aux nouvelles contingences. En absence d'OFC fonctionnel, cette adaptation est altérée, mais peut néanmoins être compensée par un apprentissage intensif (Sharpe et al., 2019; Stalnaker et al., 2015). Ces déficits ont été rapprochés de plusieurs études suggérant que l'OFC encode la valeur d'une option (O'Neill and Schultz, 2010; Padoa-Schioppa, 2011; Plassmann et al., 2007). En effet, des expériences d'imagerie fonctionnelle chez l'Homme ou des enregistrements extracellulaires chez le primate indiquent que l'activité de l'OFC reflète dans une tâche de prise de décision la valeur des récompenses ou d'autres paramètres tels que le délai ou le risque associés (Miller and Wallis, 2013; O'Neill and Schultz, 2010; Padoa-Schioppa, 2009; Plassmann et al., 2007; Rushworth et al., 2012) (Figure 11A). La représentation de valeur dans l'OFC semble par ailleurs respecter les règles d'invariance et de transitivité de la théorie neuroéconomique (Padoa-Schioppa, 2011, 2009) (cf. Chapitre 1 – III)). Ces auteurs considèrent ainsi que l'OFC encode la valeur économique d'un choix dans l'espace des offres et des buts possibles (Padoa-Schioppa, 2011). Néanmoins, l'inactivation spécifique de l'OFC latéral (lOFC) par optogénétique n'affecte pas les comportements de choix chez le rat, dans des tâches de choix binaires (Gardner et al., 2017; St. Onge and Floresco, 2010). Cela suggère que l'OFC ne serait pas nécessaire en soi aux décisions basées sur la valeur (Schoenbaum et al., 2011).

L'OFC aurait plutôt un rôle dans l'assignement de crédit temporel (Gardner and Schoenbaum, 2021; Noonan et al., 2012; Rushworth et al., 2011; Stalnaker et al., 2015; Walton et al., 2010). Cette fonction serait principalement endossée par le lOFC, qui lorsqu'il est lésé chez le singe, altère l'association action-résultat d'événements récents (Rushworth et al., 2011; Walton et al., 2010). Une étude menée chez le rat montre que l'encodage du délai associé aux récompenses par les neurones de l'OFC est indépendant de la valeur de ces récompenses (Roesch et al., 2006; Rushworth and Behrens, 2008) (Figure 11B). Les neurones signalant les récompenses (solutions de sucrose) délivrées avec de courts délais ne sont pas modulés par la quantité de ces récompenses. Ces neurones n'encodent donc pas une valeur, ni objective, ni subjective, mais une caractéristique particulière de la récompense. Plus qu'encoder une valeur, certains auteurs postulent que l'OFC signalerait plutôt des inférences, de valeur par exemple, par opposition avec les valeurs apprises par expérience (Gardner and Schoenbaum, 2021; Izquierdo et al., 2004; Stalnaker et al., 2015). Cette fonction de l'OFC, et en particulier du lOFC, l'implique davantage dans l'apprentissage et la mise à jour de nouvelles associations (Constantinople et al., 2019; Rangel et al., 2008; Rushworth et al., 2011; Sharpe et al., 2019; Sul et al., 2010; Takahashi et al., 2017).

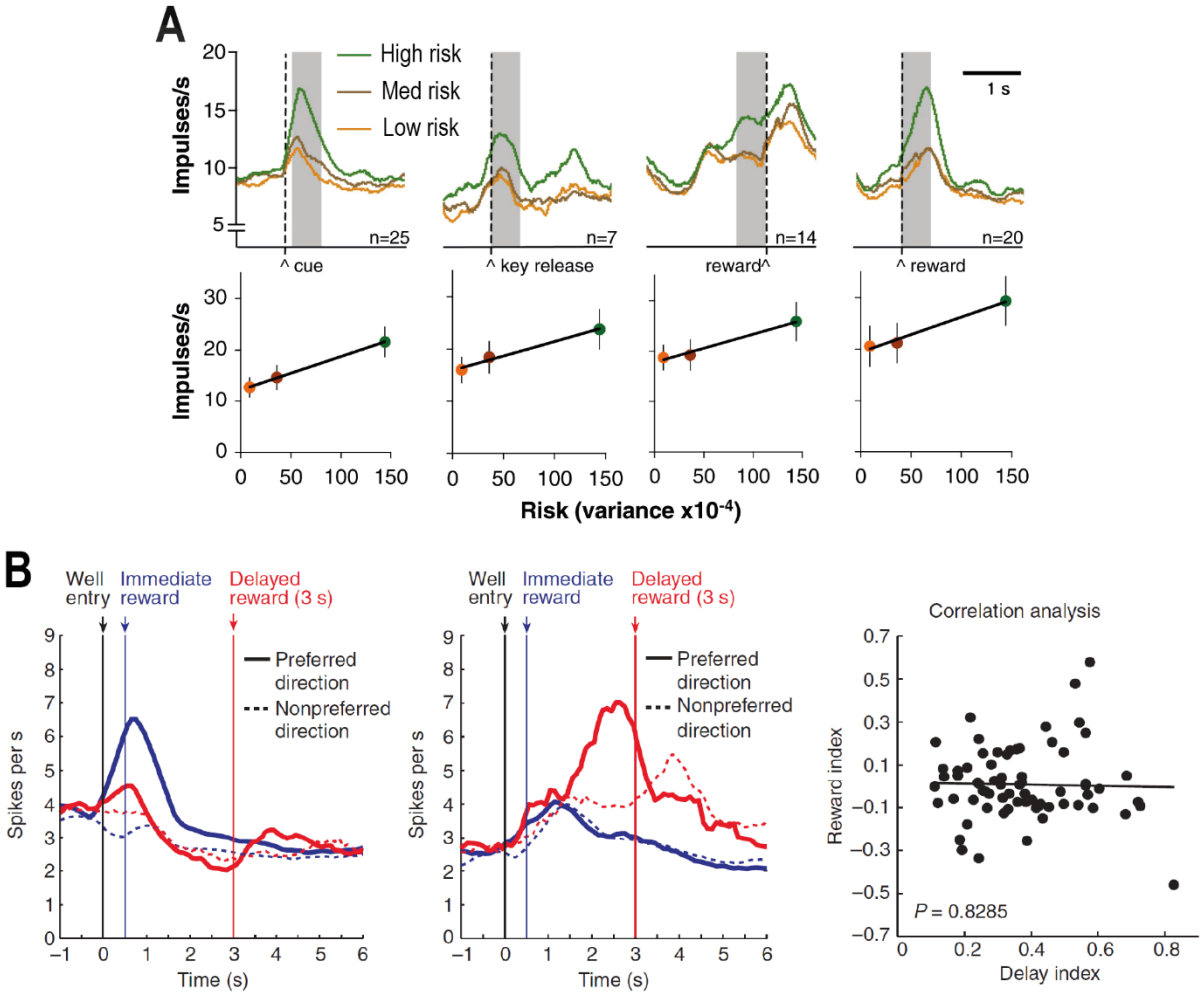


FIGURE 11 : Plusieurs encodages dans l'OFC

A. Activité moyenne de neurones d'OFC chez le singe durant une tâche de choix binaire sous risque dans laquelle l'animal peut choisir entre une option risquée (risque élevé, moyen ou faible) ou une option certaine. Une fois les deux indices présentés (*cue*), l'animal doit effectuer une saccade oculaire et relâcher la manette (*key release*) pour confirmer son choix. Haut : Les neurones présentés sont ceux dont la réponse à différentes périodes de la tâche de choix (présentation de l'indice, choix, récompense) sont positivement corrélés au risque associé à l'option choisie. Bas : Fréquence de décharge moyenne des neurones (± SEM) durant les périodes grisées sur les panels du haut. Ces neurones de l'OFC encodent le risque associé au choix à différentes étapes de la tâche. *Adapté de O'Neill and Schultz 2010.*

B. Activité moyenne de neurones d'OFC chez le rat dans une tâche de choix forcé (une seule option disponible). Gauche : Histogramme de population représentant l'activité de neurones répondant aux récompenses délivrées avec de courts délais. Milieu : Histogramme de population représentant l'activité de neurones répondant aux récompenses délivrées avec de longs délais. Les neurones présentent également une activité modulée préférentiellement dans une direction. Droite : Relation entre l'activité des neurones répondant pour des délais courts et la quantité de récompense lors de choix libres. Index de délai : *Delay Index* = (S-L)/(S+L), S et L étant l'activité en fréquence lors d'essais avec délais courts (S) ou longs (L). Index de récompense : *Reward Delay* = (B-S)/(B+S), B et S étant l'activité en fréquence lors d'essais avec grandes (B) ou petites (S) récompenses. Les neurones de l'OFC encodent le délai de manière indépendante de la valeur de la récompense. *Adapté de Rushworth & Behrens 2008 et Roesch 2006.*

Une théorie sur la fonction de l'OFC permet d'expliquer les différentes observations décrites précédemment (Wilson et al., 2014b). Selon cette théorie, l'OFC représenterait l'espace des états, c'est-à-dire l'état interne de l'individu dans l'espace des états de l'environnement à chaque instant. L'état de l'environnement correspondant à de multiples aspects d'une tâche (sa nature, sa structure, ses règles), différentes informations liées aux récompenses peuvent être signalées par les neurones de l'OFC. La notion d'état dans cette théorie englobe également les états de l'environnement qui ne sont pas explicitement indiqués par des stimuli (par exemple le délai), on parle alors d' « états cachés ». L'OFC créerait en quelque sorte une tâche cognitive de la tâche, et signalerait un message du type « vous êtes ici », et non « allez par-là », ce qui explique son rôle dans l'apprentissage de nouvelles contingences, et pourquoi son inactivation ne perturbe pas directement les choix basés sur la valeur (Sharpe et al., 2019; Stalnaker et al., 2015).

En accord avec cette théorie, il a par exemple été montré que les neurones de l'OFC chez le singe encodaient indépendamment plusieurs paramètres d'une tâche de prise de décision (Blanchard et al., 2015). Dans cette étude, les neurones de l'OFC signalent à la fois la quantité de récompense ainsi que la quantité d'information contenue dans un indice prédicteur, et ce de manière non corrélée. Les résultats obtenus chez le rat et illustrés en Figure 11B sont également en accord avec cette théorie. Cela soutient l'idée d'une représentation abstraite d'états de l'environnement plutôt que de l'intégration spécifique de la valeur, même subjective.

Cette vision pose donc la question du rôle de l'OFC dans les comportements auto-générés, qui ne sont pas spécifiquement associés à des indices déclencheurs ni même à des états cachés de l'environnement, mais à des signaux internes.

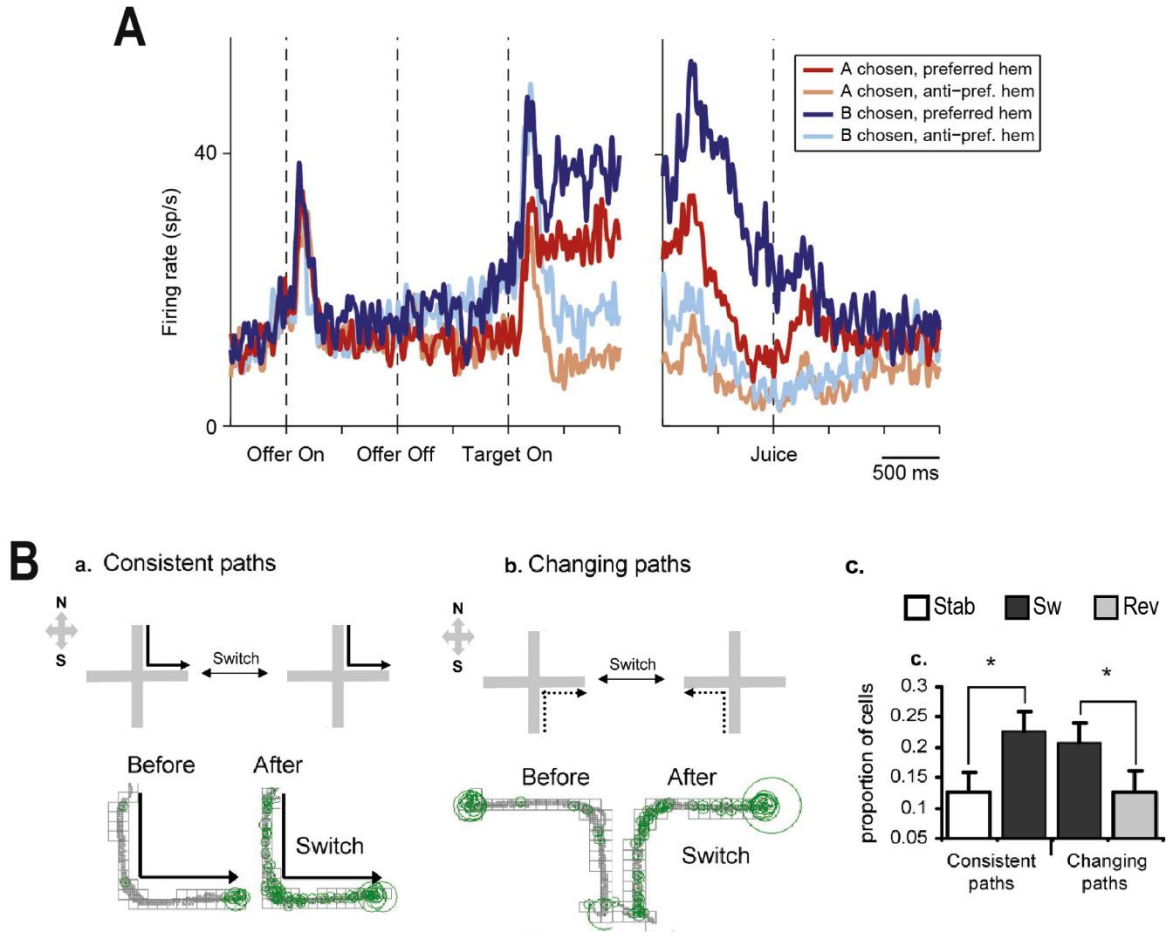
4) Le Cortex Préfrontal

Les études de lésions ou inactivations du PL PFC indiquent des altérations de performances des animaux dans des tâches de prise de décision, et notamment de flexibilité dans les choix (Miller and Wallis, 2013; Rushworth et al., 2011; Sharpe et al., 2019). Chez le rat, l'inactivation du PL PFC par infusion de muscimol (agoniste des récepteurs GABA de type A) augmente les choix associés à des options risquées, mais uniquement si les contingences de la tâche sont changeantes, indiquant une altération de flexibilité dans les stratégies de choix mises en œuvre par les rats (St. Onge and Floresco, 2010). Il a été proposé que le mPFC, et en

particulier la région PL, serait impliqué dans l'adaptation des comportements dirigés vers un but en fonction du contexte (Daw et al., 2005). Néanmoins, son rôle précis n'est pas clair car ce cortex semble être peu impliqué dans l'acquisition du comportement et dépendre des paradigmes utilisés (Miller and Wallis, 2013; Sharpe et al., 2019). L'équipe de Cohen propose que le PFC exerce un contrôle « top-down » sur les comportements en modulant les associations stimulus-réponse selon le contexte (Miller and Cohen, 2001; Sharpe et al., 2019). Il utiliserait les informations de l'environnement pour activer des chemins d'associations spécifiques, ce qui permettrait l'intégration d'informations complexes (d'ordre supérieur aux informations primaires sensori-motrices par exemple).

Dans le cadre de la prise de décision, on attribue également au PFC un rôle dans la comparaison des valeurs ou des utilités (Hunt and Hayden, 2017; Rushworth et al., 2011). En effet, chez l'Homme, des études menées en imagerie fonctionnelle indiquent que l'activité du mPFC au moment du choix corrèle et prédit la valeur de l'option (Euston et al., 2012; Rushworth et al., 2011; Rushworth and Behrens, 2008). Plus spécifiquement, le PFC ventro-médian chez l'Homme semble encoder des valeurs subjectives déduites des préférences des sujets (Cisek, 2012; Kable and Glimcher, 2007; Padoa-Schioppa, 2011; Rushworth and Behrens, 2008), mais également refléter les valeurs d'options non choisies (Lebreton et al., 2009). Chez le singe, les neurones du lPFC prédisent la valeur de l'option choisie et leur activité est modulée selon le plan d'action qui sera engagé, suggérant leur rôle dans la transformation d'une valeur économique en plan d'action (Cai and Padoa-Schioppa, 2014) (Figure 12A). Cela a conduit à postuler que le mPFC encoderait les actions possibles (les offres) dans l'environnement en fonction du but (ou de l'état central motivationnel, cf. Chapitre 1) actuel de l'animal (Cisek, 2007).

En termes computationnels, certains auteurs proposent que le mPFC utiliserait la carte cognitive des états de l'OFC, pour mettre à jour des représentations de contextes, événements, buts et actions appropriées, et contrôlerait ainsi les comportements sur un mode de fonctionnement *model-based* (Sharpe et al., 2019; Wise, 2008) (cf. Chapitre 1). Cela expliquerait l'implication du mPFC dans la planification d'actions dans l'arbre des options possibles (Dolan and Dayan, 2013). Cependant, les comportements *model-based* sont souvent testés dans des contextes probabilistes, associés à de l'incertitude sur les « branches » de l'arbre des options. On peut donc se demander à quel point le mPFC est lié à la recherche dans l'arbre en soi, ou à la résolution de l'incertitude liée à cette forme de planification (Bach and Dolan, 2012).



<u>FIGURE 12</u> : Encodages dans le PFC et changement de stratégie

A. Exemple de l'activité d'un neurone du PFC latéral ventral chez le singe lors d'une tâche de prise de décision. L'animal a le choix entre deux récompenses (jus de fruits) de valeurs différentes (type et quantité), présentées par deux indices prédicteurs (*offer on*), et doit valider son choix en effectuant une saccade oculaire au moment demandé (*target on*). Les essais sont divisés en 4 cas, selon que l'animal a choisi l'option A ou B et selon l'hémichamp visuel préféré de la cellule. Ce neurone répond à la présentation des indices prédicteurs. Son activité est modulée par le choix du jus (A ou B) avant que les cibles ne soient présentées. Après la présentation des cibles, l'activité de la cellule encode également la direction de la saccade oculaire, c'est-à-dire le plan d'action du choix. Cet exemple illustre que les neurones du PFC multiplexent différents signaux lors du processus de décision. *Adapté de Cai and Padoa-Schioppa 2014.*

B. Activité des neurones du mPFC de rats dans une tâche spatiale de changement de règle. Dans un labyrinthe surélevé en forme de « plus », les rats apprennent à obtenir des récompenses dans un paradigme de place en suivant la règle « aller à l'est », et un paradigme égocentrique en suivant la règle « tourner à gauche », ainsi qu'à s'adapter à des changements entre ces deux paradigmes. Après un changement (*switch*), le nouveau chemin emprunté peut être similaire (*consistent*, a) ou différent (*changing*, b) du précédent. **a,b** : La position de l'animal est représentée en gris, l'activité neuronale par des cercles verts dont le rayon est proportionnel au nombre de potentiels d'action dans chaque aire du labyrinthe. **c** : Proportion de neurones dont l'activité est modulée par le changement de règle (*switch*, Sw). Les conditions stables (*Stab*) et inverse (*Rev*) servent de conditions contrôle. Les neurones du mPFC répondent de manière dynamique aux changements de règles. *Adapté de Rich & Shapiro 2009.*

Un dernier rôle prêté au PL PFC est qu'il déterminerait la politique de l'animal (les règles selon lesquelles effectuer les choix, cf. Chapitre 1) plutôt que les choix directement. La lésion de cette structure chez la souris provoque des déficits lorsqu'il s'agit d'alterner entre un comportement opérant pour récupérer une récompense et une stratégie d'exploration face à de la nouveauté (Granon and Changeux, 2012). Chez le rat, les neurones du mPFC semblent s'activer spécifiquement après des changements de règles dans un même contexte (Rich and Shapiro, 2009) (Figure 12B). Par ailleurs, certains auteurs proposent que le PFC serait responsable du désengagement vis-à-vis d'une stratégie habituelle au profit d'une stratégie plus flexible et potentiellement plus exploratoire (Cohen et al., 2007; Daw et al., 2005). Des modèles d'apprentissage par renforcement inspirés des propriétés neuroanatomiques du PFC indiquent qu'il pourrait réguler les paramètres de politique au cours d'une tâche (Khamassi et al., 2011).

Les différents rôles suggérés du PFC dans la mémoire de travail et la flexibilité comportementale amènent à se questionner sur son rôle spécifique dans la prise de décision. Est-il effectivement le lieu de calculs de comparaison de valeurs, ou plutôt l'instigateur des changements de stratégies utilisées pour prendre des décisions ? Si le PL PFC est vu comme le chef d'orchestre des comportements *model based*, est-il également un acteur clé des décisions auto-générées (en général, ou en condition incertaine) ?

Une première vision des choses a été d'attribuer à chaque région préfrontale une fonction spécifique. Au même titre que les cortex somato-sensoriels représentent de l'information à propos de chaque partie du corps, certains auteurs postulent que les cortex préfrontaux encoderaient, représenteraient et stockeraient de l'information sur les comportements, ainsi que leurs conséquences probables en fonction du contexte (Euston et al., 2012; Wise, 2008). Mais, comme évoqué précédemment, il semble que plusieurs aires préfrontales partagent parfois le même rôle. Un autre point de vue est donc de considérer que les circuits préfrontaux dans la prise de décision peuvent participer aux mêmes computations, de façon distribuée et en parallèle (Cisek, 2012; Hunt and Hayden, 2017; Rushworth et al., 2012).

En outre, ils semblent dialoguer à plusieurs échelles temporelles avec d'autres régions cérébrales elles-mêmes impliquées dans les comportements dirigés vers les récompenses. C'est en particulier leur interaction avec la VTA qui nous intéresse dans le cadre de la prise de décision et des comportements exploratoires.

III) Interactions entre l'Aire Tegmentale Ventrale et le Cortex Orbitofrontal et Préfrontal

Il est admis que les processus de décision émergent de systèmes interagissant à large échelle et incluant de multiples composantes corticales et sous-corticales (Bechara et al., 2000; Cisek, 2012; Hunt and Hayden, 2017; Yun et al., 2020). Les cortex préfrontaux et les neurones DAergiques de la VTA jouent causalement dans différents aspects des comportements dirigés vers des récompenses et dans la prise de décision. Néanmoins, leurs mécanismes d'interactions et de contrôle des comportements auto-générés restent peu clairs.

Il semblerait que l'OFC et le PFC participent au calcul de RPE, l'OFC en encodant notamment des représentations d'état (Gershman and Uchida, 2019; Stalnaker et al., 2015), et le mPFC les prédictions de valeurs choisies (Lak et al., 2019). Ces informations seraient ensuite intégrées par les neurones DAergiques de la VTA et serviraient ainsi à l'apprentissage (Hunt and Hayden, 2017; Jo and Mizumori, 2016; Sharpe et al., 2019). En effet, la lésion ou l'inactivation de l'OFC chez le rat affecte la représentation par les neurones DAergiques de conséquences (*outcome*) inattendues et perturbe la formation de nouvelles associations (Takahashi et al., 2011, 2009). L'inhibition du mPFC chez la souris par pharmacogénétique altère le calcul de RPE par les neurones DAergiques, et ce, uniquement lorsque les états de l'environnement et des récompenses sont variables (états cachés, c'est-à-dire non spécifiés par un stimulus) et que leur valeur repose sur des inférences (Starkweather et al., 2018).

Des études de pharmacologie des récepteurs DAergiques dans le PFC, et particulièrement dans la région prélimbique, suggèrent que la DA dans le PFC est impliquée dans la flexibilité comportementale (Floresco, 2013; Verharen et al., 2018). La manipulation des niveaux de DA dans le PL PFC par infusion locale chez le rat d'agonistes ou antagonistes des récepteurs DAergiques a permis de relier une élévation de la transmission DAergique avec des déficits de discrimination d'indices, et une diminution de cette transmission avec une diminution de choix risqués dans une tâche de décision (Floresco, 2013). Par ailleurs, tandis que la stimulation phasique des terminaisons DAergiques dans le mPFC chez la souris ne semble pas induire de renforcement, leur stimulation tonique ou phasique semble altérer le degré de flexibilité dans les choix (Ellwood et al., 2017). Ces expériences suggèrent une implication de la DA préfrontale dans la politique et/ou l'exploration, mais ne permettent néanmoins pas de conclure spécifiquement sur le rôle de l'activité tonique ou phasique de la DA, ni comment elle affecte l'activité des cortex préfrontaux.

Ceci dit, d'autres travaux permettent de proposer que la DA affecte la dynamique des réseaux corticaux et en particulier des oscillations corticales. L'inhibition pharmacologique des récepteurs DAergiques dans le PFC chez le singe augmente la puissance des oscillations α (8-14 Hz) et β (14-30 Hz) dans une tâche de conditionnement et de prise de décision (Puig et al., 2014). Chez le rat, l'infusion locale de DA dans le PFC induit une augmentation de cohérence dans la gamme θ entre le PFC et une aire distante, l'HPC (Benchenane et al., 2010). Le PFC et la VTA se synchroniserait via une oscillation à 4Hz transitoire au niveau de la zone de choix dans une tâche de labyrinthe en Y (Fujisawa and Buzsáki, 2011), ou via des oscillations θ dans des contextes de choix (Park and Moghaddam, 2017). Par ailleurs, l'activité des neurones DAergiques de la VTA semble moduler directement le couplage θ – γ des EFP dans le mPFC chez le rat (Lohani et al., 2019). Un des mécanismes d'action de la DA dans le PFC propose que la DA favoriserait la transition des neurones du PFC d'un état « bas » hyperpolarisé (*down state*) à un état « haut » dépolarisé (*up state*) (Seamans and Yang, 2004). Par ailleurs, la fonction modulatrice de la DA et sa capacité à agir à large échelle sur une population de neurones faciliterait la synchronisation au niveau du réseau. Dans le cadre de la mémoire de travail, la transmission DAergique par le récepteur D1 a été associée à la stabilité des représentations au sein d'assemblées neuronales, tandis que celle par le récepteur D2 à des représentations moins stables et potentiellement plus flexibles (Durstewitz and Seamans, 2008; Puig et al., 2014). On peut donc se demander comment la DA participe aux transitions entre assemblées préfrontales représentant des choix ou des politiques.

Dans la boucle-mésocorticale reliant les cortex OFC et PFC et la VTA, les informations émergeant des cortex semblent liées à des stimuli ou à des états et influencer l'activité phasique des neurones DAergiques et donc le signal de RPE. Au contraire, l'influence de la DA sur les cortex semble impliquer plutôt le tonus DAergique et influer notamment sur la recherche du risque. Les liens entre ces structures au moment de la décision et de l'action restent en revanche mal connus. En effet, si l'OFC signale des états internes cachés ou inférés, le lien entre ces représentations et le calcul de RPE par les neurones DAergiques reste encore à déterminer. De même, les mécanismes par lesquels l'encodage des valeurs - sous forme de monnaie commune - par les neurones DAergiques (cf. Chapitre 2) affecte le contrôle du PFC sur les stratégies comportementales des animaux sont mal compris. En particulier lorsque le comportement est auto-généré, comment L'OFC, le PFC et la VTA interagissent au moment du choix et contrôle l'initiation du comportement dirigé ?

Chapitre 4 : Paradigme comportemental

Au cours de ma thèse, j'ai utilisé une tâche basée sur un paradigme d'apprentissage par renforcement développé au laboratoire. Cette tâche permet de conditionner des souris à effectuer des choix successifs entre deux options. Il s'agit d'une adaptation pour la souris (Naudé et al., 2016) d'une tâche classique dite de "bandit à bras multiples" (Cohen et al., 2007).

I) La tâche 3 points

1) Principe de la tâche

Grâce à un jeu de plusieurs emplacements marqués au sol dans une arène circulaire - on parle également de cibles ou de points – et associés à des stimulations intra-crâniales au niveau du MFB (faisceau d'axones dopaminergiques issus, entre autres, de la VTA et dont la stimulation mime l'effet d'une récompense, cf. Chapitre 2 – II) – 1)), les souris apprennent à réaliser une succession de choix binaires entre deux alternatives (Figure 13A). Si l'on fait en sorte que l'animal ne puisse pas obtenir deux stimulations consécutives au même point, trois cibles sont suffisantes pour permettre d'enchainer des successions de choix binaires. Ainsi, lorsqu'un animal arrive sur une cible, il reçoit une stimulation intra-crâniale, et ne pouvant pas en obtenir une seconde sur cette même cible, il doit se diriger vers l'une ou l'autre des deux cibles restantes pour obtenir une nouvelle stimulation. Cela constitue ainsi une séquence de choix binaires. Lorsqu'une souris est détectée dans l'une des trois zones de récompense, un stimulateur électrique délivre une stimulation au niveau du MFB (train de 200 ms d'ondes carrées biphasiques de 5 ms pulsées à 100 Hz (20 impulsions par train)). L'intensité est ajustée pour chaque souris dans une fourchette de 10 à 200 μA pendant l'entraînement, puis maintenue constante pendant toute l'expérience. L'ajustement est fait de sorte que les souris atteignent un total de 50 à 100 récompenses et le stabilisent pendant deux sessions (de 5min) successives.

2) Variations de la tâche

En partant de ce principe de base, qui allie une autostimulation intracrânienne (ICSS, *intracranial self-stimulation*) du MFB unique en un point, et l'alternance entre trois points marqués au sol, plusieurs règles de délivrance de récompenses peuvent être mises en place afin d'étudier différents aspects de la prise de décision et du choix :

• Cadre déterministe ou certain : L'ICSS est délivrée de manière certaine en chaque cible (Figure 13A). C'est ce cadre qui est généralement utilisée pour l'apprentissage de la tâche. L'animal qui n'a encore jamais été exposé aux stimulations électriques du MFB dans cet environnement apprend à faire l'association entre le marquage au sol et la stimulation. Après quelques sessions, l'animal enchaine des séquences de choix. Dans ce cadre déterministe, la souris doit en chaque point faire le choix entre deux options de même récompense.

• Cadre probabiliste ou incertain : Dans ce cadre, chaque cible est associée à une probabilité différente de délivrer la stimulation (Figure 13A). Le jeu de probabilités classiquement utilisé au laboratoire est 25%, 50% et 100%. Puisque les souris ne peuvent pas obtenir deux récompenses consécutives sur la même cible, elles sont soumises à trois choix binaires distincts (« paris » ou *Gambles*, G), notés respectivement G_{25}, G_{50} et G_{100} selon la cible sur laquelle elles se trouvent au moment du choix (Figure 13A). Chacun de ces paris implique un choix entre deux gains, ici : G_{25} = {p_{100} contre p_{50}}, G_{50} = {p_{100} contre p_{25}} et G_{100} = {p_{50} contre p_{25}}. Manipuler les probabilités dans une série de choix binaires permet de démêler les influences de deux paramètres sur les choix des animaux : i) le gain espéré en chaque point (c'est-à-dire la probabilité, ici 0.25, 0.5 ou 1), et ii) l'incertitude associée qui correspond à la variance de la probabilité, calculée par la formule $p(1-p)$ (ici 0.1875, 0.25 et 0 respectivement). Dans cette tâche, la distribution de la récompense est un événement binaire et suit donc une loi binomiale. Cette règle de distribution de récompense permet d'étudier les décisions des animaux sous incertitude.

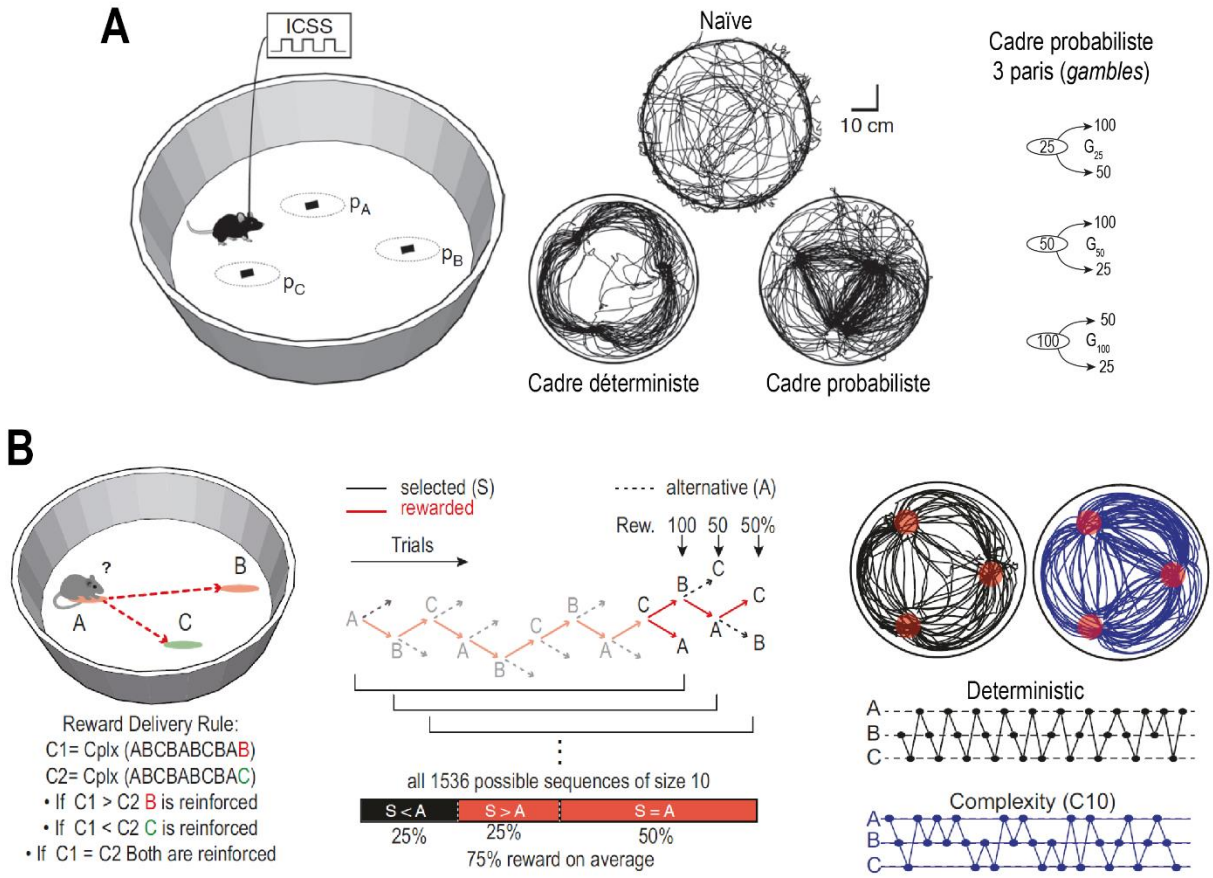


FIGURE 13 : Les différentes variantes de la tâche 3 points

A. Tâche 3 points en cadres déterministe et probabiliste. Gauche : Schéma de la tâche. Trois cibles sont explicitement marquées au sol, formant un triangle équilatéral de 50 cm de côté, dans une arène de 80 cm de diamètre. En chaque cible, une autostimulation intracrânienne (ICSS) est délivrée avec une probabilité p (100% en chaque point en cadre déterministe ; 25%, 50% et 100% en cadre probabiliste). Milieu : Exemples de trajectoires typiques d'une souris avant l'apprentissage (haut), à la fin du cadre déterministe (bas) et à la fin du cadre probabiliste (droite). Droite : Schéma des trois différents paris (*Gambles*) auxquels les souris sont soumises en cadre probabiliste. *Adapté de Naudé 2016.*

B. Tâche 3 points en cadre complexe. Gauche : Schéma de la tâche en cadre complexe et de la règle de distribution des récompenses utilisé par l'algorithme. A chaque essai, la cible renforcée est celle maximisant la complexité de la chaîne de 10 choix successifs (les 9 derniers choix et le suivant). Milieu : Illustration de l'arbre de distribution des récompenses dans la tâche. En théorie, si l'animal se déplace au hasard dans cet arbre, il obtiendra un taux de succès moyen de 75% à l'issu de la session. Droite : Exemples de trajectoires typiques d'une souris à la fin du cadre déterministe (noir) et du cadre complexe (bleu). *Adapté de Belkaid 2020.*

• Cadre complexe : Dans ce cadre, la distribution de l'ICSS est associée à un algorithme qui estime, en ligne, à chaque essai, la complexité grammaticale de Lempel-Ziv (Lempel and Ziv, 1976) des deux séquences constituées des 9 choix précédents et du prochain choix parmi les deux cibles disponibles (Figure 13B). L'option récompensée est celle correspondant à la séquence de plus grande complexité. Si les complexités correspondant aux deux options sont égales, elles sont toutes deux récompensées. Une autre manière de représenter cette tâche est de redéployer les séquences de choix au sol, les transformant en chemins dans un arbre de choix binaires. L'algorithme de complexité revient alors à distribuer les récompenses de manière déterministe dans cet arbre : en effet il est possible, à chaque essai, de savoir si telle option sera récompensée ou non, la distribution n'étant pas stochastique. Cette distribution bien que déterministe a ceci de particulier i) qu'un agent peut a priori apprendre l'ensemble des combinaisons gagnantes ou apprendre à répéter une combinaison gagnante et donc avoir un taux de succès de 100%, ii) qu'un agent n'effectuant ses choix qu'au hasard (en tirant à pile ou face à chaque choix) atteindra quand même un taux de succès moyen de 75% de récompenses reçues et enfin iii) qu'un agent répétant une même séquence (associée donc à une faible complexité, par exemple tourner en rond A-B-C-A) perdra à chaque coup. Cette règle de distribution de récompense permet d'étudier la variabilité (ou l'aléatoire) dans les décisions.

II) Comportement des souris dans la tâche

1) L'essai comme comportement de base de la tâche

Une fois la tâche apprise, les comportements de choix, performances et trajectoires des animaux atteignent un équilibre, et ces derniers se comportent dans la tâche de la même manière en moyenne d'un jour à l'autre. Ils ont appris l'environnement statistique et les contingences qui amènent à la délivrance d'une récompense, et maintiennent ensuite un comportement stable. Une session (d'une durée de 5 minutes) se compose alors d'une succession d'essais amenant une souris d'une cible sur une autre. Ces essais constituent l'unité d'analyse de comportement de base dans la tâche. Un essai est caractérisé par l'initiation d'un mouvement d'une cible donnée vers la cible suivante, suivi d'un ralentissement au moment de l'arrivée sur cette cible

et d'une pause avant la reprise d'un cycle. Ces essais sont auto-générés dans le sens où ils ne sont pas déclenchés par un stimulus extérieur mais relèvent d'une motivation intrinsèque.

On peut observer une importante variabilité d'un essai à l'autre. Tout d'abord, il y a une variabilité inter-essais substantielle i) dans le temps durant lequel les animaux restent à l'arrêt (au niveau d'une cible) avant d'initier un nouvel essai, ii) dans le temps pour atteindre la vitesse maximale, et iii) dans le temps global pour aller d'une cible à une autre. Ensuite, sur chaque cible, les souris peuvent soit continuer le mouvement de l'essai précèdent en avant le long des trois emplacements et donc faire des séquences circulaires (c'est-à-dire de type A-B-C si ces lettres désignent chacune des cibles), soit faire demi-tour en revenant en arrière, à l'essai i+1, vers l'emplacement i-1 et donc faire une séquence de type A-B-A. Nous verrons dans l'article 1 que la direction (demi-tour ou marche avant) a un impact sur la durée de l'arrêt (qui est un moment précoce de l'essai), et enfin sur le temps nécessaire pour atteindre le but. Cette addition successive de variabilité à chaque étape d'un essai suggère que la direction de la trajectoire, l'initiation et la vigueur ne sont pas déterminés par l'environnement mais que nous avons à faire à des séquences comportementales auto-générés (cf. Article 1) avec plusieurs étapes et décisions successives : la direction ou le choix de la prochaine cible, mais aussi à quel moment partir, l'organisation de la séquence motrice, etc.

2) Comportement locomoteur dans la tâche

Sur la base de la séquence comportementale précédemment décrite, les souris développent des trajectoires et comportements locomoteurs adaptés aux contextes de la tâche auxquels elles sont confrontées (cadre déterministe, probabiliste ou complexe). Je vais décrire quelques effets de la tâche sur la locomotion des animaux, déjà publiés dans un article du laboratoire (Naudé et al., 2016).

Tout d'abord, dans la tâche déterministe, la vitesse maximale des souris est affectée par l'intensité de l'ICSS (Figure 14A). Après apprentissage, si on fait varier l'intensité du courant délivré, on observe que plus l'intensité est forte (c'est-à-dire plus la magnitude de la récompense est grande), plus la vitesse maximale est élevée. Cela indique que l'intensité de la récompense affecte la vigueur des actions dans la tâche. Le temps auquel la vitesse maximale est atteinte dans un essai n'est en revanche pas impacté. Au contraire, les probabilités de

récompense associées à chaque cible en cadre probabiliste n'affectent pas la valeur de la vitesse maximale dans l'essai (Figure 14B).

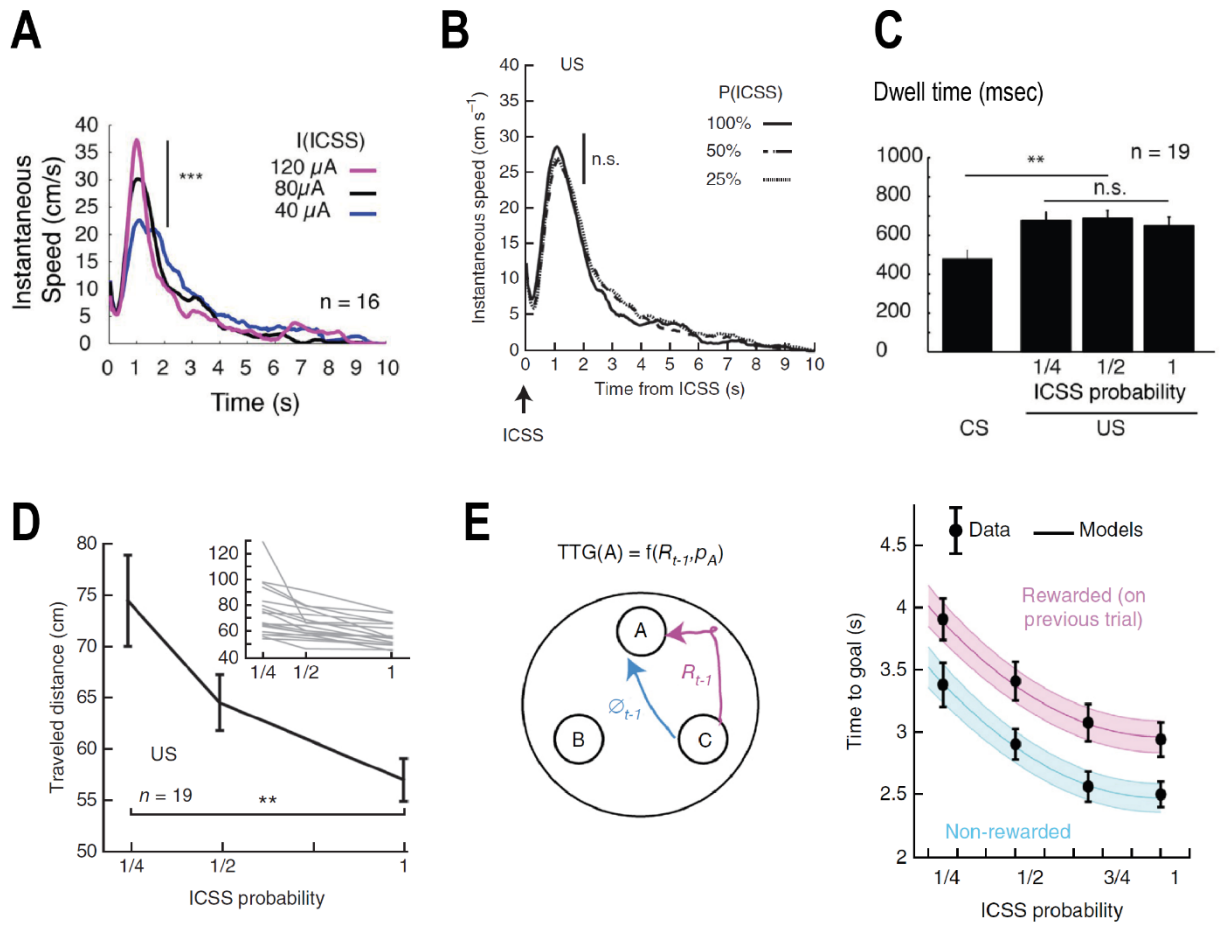


FIGURE 14 : Comportement locomoteur dans la tâche 3 points

A. Effet de l'intensité de l'ICSS sur la vitesse des animaux en cadre déterministe. Les profils de vitesse instantanée sont moyennés après apprentissage. La vitesse maximale au sein des essais augmente avec l'intensité de la stimulation délivrée.

B. Effet des probabilités sur la vitesse des animaux en cadre probabiliste. Pour chaque profil moyen de vitesse instantanée, les essais sont triés selon la probabilité de récompense associée au point vers lequel se dirigent les animaux. La vitesse n'est pas affectée par les probabilités.

C. Effet des probabilités sur le temps de résidence en chaque point (*dwell time*). A la fin du cadre probabiliste, le temps de résidence est plus élevé qu'à la fin du cadre déterministe.

D. Effet des probabilités sur la distance parcourue entre deux cibles à la fin du cadre probabiliste. Les animaux parcourent de plus grandes distances lorsqu'ils se dirigent vers des récompenses associées à de plus faibles probabilités.

E. Effet de l'omission sur la durée des essais à la fin du cadre probabiliste. Gauche : Schéma du modèle de locomotion dans la tâche. Droite : Durée des essais en fonction de la probabilité associée au point vers lequel se dirigent les animaux, dans le cas où l'essai précédent a été récompensé (violet) ou non (bleu).

Adapté de Naudé 2016.

Ensuite, le temps de résidence, qui correspond à la durée durant laquelle l'animal s'arrête sur un point après l'avoir atteint (*dwell time*), augmente en cadre probabiliste par rapport au cadre déterministe (Figure 14C). Cette durée n'est cependant pas impactée par la valeur de la probabilité de récompense associée au point vers lequel l'animal se dirigera. En revanche, les probabilités influencent la distance parcourue entre deux points (Figure 14D). En effet, les animaux parcourent de plus grandes distances lorsqu'ils se dirigent vers des récompenses associées à de plus faibles probabilités. Cela reflète leur tendance à explorer le reste de l'arène entre deux cibles, surtout lorsque la prochaine récompense est peu probable. Cela explique également que la durée de l'essai dépende de la probabilité de la cible choisie par l'animal : si celui-ci parcourt une plus grande distance en se dirigeant vers la cible associée à la probabilité 25%, il met également plus de temps pour y parvenir (Figure 14E).

Enfin, si la probabilité affecte la distance parcourue (Figure 14D) et pas la vitesse (Figure 14B), l'historique de récompense du choix précédent impacte également la durée des essais (Figure 14E). Lorsque le choix précédent s'est soldé par une omission, la durée de l'essai suivant est plus courte. Un modèle de locomotion intégrant à la fois l'historique du choix précédent (en termes de récompense) et la probabilité de récompense de l'option choisie, permet de reproduire ces données. Cela suggère que l'incertitude attendue influence à la fois la tendance à explorer, mais aussi la motivation pour atteindre le but de manière directe.

3) Incertitude et balance Exploration / Exploitation

Parmi les composantes de la prise de décision, le compromis (ou la balance) entre exploration et exploitation présente un intérêt majeur. L'exploitation consiste à choisir l'option qui, sur la base de l'historique des récompenses reçues par l'animal, semble optimale. Cependant, confrontés à un choix binaire entre deux alternatives, l'une avec une faible probabilité de récompense et l'autre avec une forte probabilité, les animaux ne se contentent pas d'exploiter : ils choisissent également l'option la moins susceptible d'être récompensée une grande partie du temps. L'origine de ces choix au premier abord sous-optimaux reste mal comprise. Nous l'avons interprété dans nos tâches comme de l'exploration.

L'exploration est essentielle pour recueillir des informations sur des résultats inconnus ou incertains dans un environnement changeant. C'est une stratégie adaptative qui consiste,

lorsqu'on ne connaît pas la valeur d'une récompense, à la tester pour obtenir des informations sur cette valeur, qui pourrait potentiellement être plus élevée que celle associée à la stratégie comportementale dominante jusqu'à lors. L'exploration permet donc de réduire l'incertitude sur la valeur associée à une récompense. Dans notre expérience, dans le cadre déterministe (Figure 13A), il n'y a pas d'incertitude associée aux différentes récompenses. A valeur d'option égale (aux aspects moteurs près), la stratégie de choix dominante de l'animal est circulaire, il utilise néanmoins des demi-tours d'une manière qui semble aléatoire. Nous l'interprétons comme de la variabilité dans le choix, une forme primaire d'exploration.

Dans le cadre probabiliste, l'animal est confronté à des probabilités fixes de récompense (classiquement p = 0.25, 0.5 et 1, Figure 13A), c'est-à-dire qu'il n'y a pas de changement dans les règles de distribution des récompenses au cours de l'expérience. L'incertitude dans la tâche a donc deux sources : la qualité de l'estimation par l'animal des valeurs de récompense (valeur subjective, qui peut être rapprochée de l'espérance mathématique), et la variance liée aux probabilités de récompense (maximale au point 0.5 et nulle au point 1). La souris peut diminuer l'incertitude de son estimation en acquérant davantage d'expérience au cours de l'apprentissage par exemple. C'est de l'exploration classique. Enfin, il y a une incertitude incompressible, sur laquelle l'animal n'a pas d'influence, et décrite par la variance associée à chaque probabilité. Cette variance est plus forte sur le point associé à 50% de récompense que sur le point associé à 25% et que sur le point associé à 100% où elle est nulle (Figure 15Aa). Cette incertitude mesure l'imprévisibilité du résultat. La question de savoir si les animaux en général, et les rongeurs en particulier, i) font une distinction entre ces deux formes d'incertitudes et ii) si elles sont séparables, est ouverte au débat. Même si un animal peut avoir une bonne estimation des valeurs associées à chaque récompense (la probabilité), son incompréhension du mécanisme ou de la notion de « probabilités » fait qu'il peut rester surpris et déstabilisé par l'absence au coup par coup d'une récompense (omission). Ces formes d'incertitudes génèrent de l'exploration et un précèdent travail du laboratoire propose que l'incertitude attendue (celle associé à la variance) peut se traduire par un bonus de valeur qui s'additionne à la valeur de la récompense (probabilité).

Enfin dans la troisième expérience (cadre complexe, Figure 13B), l'animal est confronté à un paradigme de choix dans lequel trois stratégies peuvent théoriquement émerger. Si l'animal produit les mêmes séquences de choix (par exemple : ABCABCABC…), il n'est pas récompensé. Si l'animal apprend à produire une ou plusieurs séquences récompensées (d'une longueur de 10) qu'il répète, il peut obtenir 100% de récompense. Enfin si l'animal choisit au hasard, il obtient 75% de récompense en moyenne. On observe que les animaux ne semblent

pas capables d'utiliser de règles particulières pour obtenir les récompenses, ni d'heuristiques dépendantes d'une mémoire de travail (en tachant d'inférer des préférences à partir des N choix précédents par exemple) : autrement dit, il n'y a pas d'exploitation. En revanche, ils sont, dans leur processus de choix, capables de générer de la variabilité et le modèle qui reproduit le mieux leurs choix est un agent aléatoire, quand bien même la règle régissant la distribution des récompenses n'est ni stochastique ni volatile. On parle d'exploration aléatoire, que l'on distingue de l'exploration dirigée vue dans le cadre probabiliste (cf. Chapitre 1 – IV)). Cette dernière vise à recueillir des informations sur les contingences environnementales spécifiquement orientées vers les options nouvelles, incertaines ou qui ont été pas ou peu visitées dans le passé, tandis que l'exploration aléatoire introduit de la variabilité indépendamment des contingences.

III) Modélisation des choix dans la tâche

1) Le modèle

Les différents modèles de compromis exploitation/exploration sont des variantes de la loi d'appariement (*Matching Law*, cf. Chapitre 1 – IV) – 2)), dans laquelle la proportion des choix dépend de la différence entre leurs valeurs attendues. Nous avons utilisé un modèle de prise de décision basé sur l'utilisation d'un Softmax (une fonction exponentielle normalisée, cf. Chapitre 1 – IV) – 2)). Cette fonction détermine la probabilité P_i de choisir la cible *i*. Dans notre tâche, puisque les souris ne peuvent pas retourner sur la même cible à l'issu d'un essai, elles doivent choisir en chaque cible entre les deux cibles restantes. En conséquence, nous avons modélisé un processus de décisions entre deux alternatives A et B et utilisé une règle de choix définie par :

$$P_A = \frac{1}{1+e^{-\beta(V_A - V_B)}} \qquad (1)$$

Où β est un paramètre de température inverse, reflétant la sensibilité du choix à la différence entre les valeurs des options, et V_i la valeur d'une option *i*.

Dans notre modèle, V_i est définie comme :

$$V_i = p_i + \varphi * p_i * (1 - p_i) - \kappa \qquad (2)$$

Où p_A désigne la probabilité de recevoir une stimulation au point A, $p_i * (1 - p_i)$ la variance (ou incertitude attendue), φ le bonus d'incertitude, et κ le coût du demi-tour.

Formellement dans le cadre probabiliste et avec $p_A > p_B > p_C$, notre modèle utilise trois équations représentant les trois choix en fonction du point de départ A, B ou C :

$$P_{A|C} = \frac{1}{1+e^{-\beta(V_A - V_B)}} \; ; P_{B|A} = \frac{1}{1+e^{-\beta(V_B - V_C)}} \; ; P_{A|B} = \frac{1}{1+e^{-\beta(V_A - V_C)}}$$

Dans ce modèle, le paramètre β reflète à quel point la différence de valeur totale entre les deux options ($\Delta V = V_A - V_B$) se traduit par une préférence pour l'option de plus haute valeur dans un pari donné en lui attribuant un poids. Si β est petit, les choix sont peu sensibles à ΔV. A la limite $\beta = 0$, les deux options ont la même probabilité d'être choisies dans chaque pari (G_{100} = G_{50} = G_{25} = 50%), ce qui conduit à une distribution égale des visites des cibles (p_{100} = p_{50} = p_{25} = 33%), indépendamment de leur valeur respective. Au contraire, une forte valeur de β indique une forte sensibilité à ΔV. Avec un $\beta \to \infty$, l'option associée à la valeur la plus importante est toujours choisie (G_{100} = G_{50} = 100% et G_{25} n'existerait pas si on considère que les animaux ne visitent jamais la cible associée à la probabilité de 25%), ce qui conduit à une distribution des visites uniquement entre les cibles associées à 100% et 50% (p_{100} = p_{50} = 50% et p_{25} = 0%). Ce paramètre β reflète donc la balance ou le compromis entre exploitation et exploration aléatoire.

Le paramètre de recherche d'incertitude φ représente le bonus de valeur accordé à l'incertitude attendue, associé à une cible. Dans notre tâche si les cibles sont associées à des probabilités de 1, 0.5 et 0.25, l'incertitude, définie formellement par la variance $p*(1-p)$, est respectivement de 0, 0.25 et 0.1875. Ce paramètre φ introduit la notion d'exploration dirigée vers l'incertitude.

Enfin, le paramètre de coût moteur κ traduit le fait que les animaux ont tendance à favoriser les séquences circulaires par rapport aux demi-tours pour minimiser les angles et optimiser une séquence motrice qui maximise les récompenses en réduisant l'effort (ici moteur). Puisqu'il représente un coût, il a une valeur positive mais est associé à une baisse de la valeur V_i (d'où le $-\kappa$ dans l'équation (2)).

2) L'estimation des paramètres du modèle

Les paramètres libres du modèle (β, φ et κ) sont ajustés en maximisant la vraisemblance des données. Étant donnée une séquence de choix $c = c_{1,\ldots,T}$, la vraisemblance des données est le produit de leur probabilité (donnée par l'équation (1)). Nous avons utilisé la fonction *optim* de R ou *fmincon* de Matlab pour effectuer les ajustements, avec les contraintes que $\beta \in]0,10]$, $\varphi \in]-1,5]$ et $\kappa \in]0,5]$. Une fois les paramètres estimés, il est possible de re-simuler des données en utilisant les modèles optimisés sur la base de ces paramètres, et de comparer les données expérimentales avec les données simulées.

Dans une première étude du laboratoire (Naudé et al., 2016), les choix des animaux en cadre probabiliste ont été modélisés à l'aide de trois modèles de prise de décision intégrant différemment l'influence de la probabilité de récompense et de l'incertitude sur les choix (Figure 15a). Dans ce cadre, les choix des souris peuvent être représentés, pour chaque pari (*gamble*), par la proportion de transitions vers le point associé à la plus haute valeur de récompense, autrement dit, par la fréquence d'exploitation (Figure 15Ab). Les choix ont été modélisés par le modèle ε-greedy, par un Softmax basé sur un paramètre (β), ou sur deux paramètres (β et φ) (Figure 15c). Le modèle ε-greedy résout le compromis exploitation / exploration à l'aide d'un paramètre ε d'exploration aléatoire. Selon ce modèle, l'option de plus grande valeur attendue est choisie avec une probabilité (1 - ε) (l'agent est « avare », *greedy*), c'est-à-dire qu'une portion ε du temps, l'autre alternative est aléatoirement explorée. Les options de moins grande valeur sont donc explorées aléatoirement et indépendamment de leurs valeurs respectives. Dans le cadre probabiliste de la tâche 3 points, ce modèle prédit un taux d'exploration identique pour les trois paris G_{25}, G_{50} et G_{100}. Le modèle Softmax basé sur le paramètre β formalise les choix selon la différence de valeur des options possibles, le paramètre β permettant de régler la sensibilité à cette différence de valeur (un grand β indiquant une tendance à maximiser le gain, c'est-à-dire à exploiter, cf. plus haut). Ce modèle prédit en cadre probabiliste une plus grande propension à exploiter l'option 100% dans le pari G_{25}, et une moins grande propension à exploiter l'option 50% dans le pari G_{100}, ce qui ne correspond pas aux comportements observés des souris (Figure 15Ab). Le modèle Softmax basé sur deux paramètres β et φ repose sur l'utilisation d'un paramètre supplémentaire ϕ qui augmente la valeur des options incertaines (de variance élevée) et permet d'expliquer des formes continues et dirigées d'exploration (cf. plus haut).

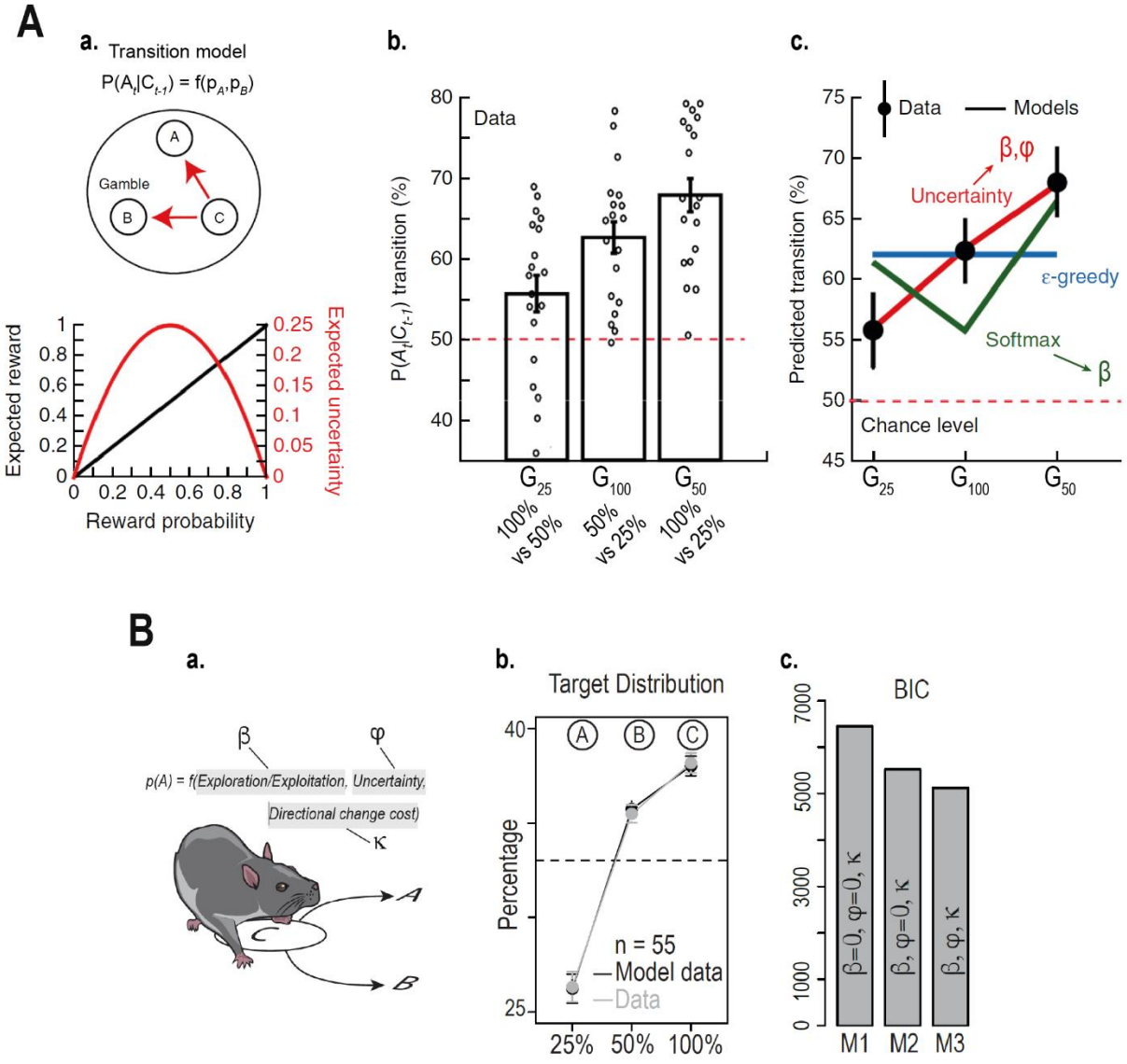


FIGURE 15 : Modélisation des décisions en cadre probabiliste

A. L'incertitude associée aux récompenses affecte les décisions des animaux. **a :** Illustration du modèle de transition. Chaque cible (A,B,C) est considérée comme un état et la probabilité de transiter d'un état à un autre dépend des probabilités de récompense des deux options disponibles. A chaque probabilité de récompense est associée une valeur attendue de récompense et d'incertitude. **b :** Proportions des choix d'exploitation (c'est-à-dire de choix associés à la cible de plus haute valeur) dans le cas des trois paris (*Gambles G1, G2, G3,* aussi appelés respectivement G_{25}, G_{100}, G_{50} dans le texte). Les points correspondent aux données individuelles de chaque animal. **c :** Transitions prédites par les modèles ε-greedy (bleu), Softmax classique (vert) intégrant seulement le paramètre β (sensibilité à la récompense) et Softmax avec bonus d'incertitude (rouge) intégrant les paramètres β et φ (bonus d'incertitude). Les données individuelles des souris sont prédites par le modèle d'incertitude. *Adapté de Naudé 2016.*

Suite de la légende page suivante

B. Intégrer un paramètre de bonus à l'incertitude ainsi que de coût de demi-tour améliore les prédictions du modèle. **a :** Principe du modèle adapté du Softmax utilisant trois paramètres β (sensibilité à la récompense ou balance exploration / exploitation), φ (bonus d'incertitude) et κ (coût lié à l'effort d'un demi-tour). **b :** Distribution des choix entre les trois points pour 55 souris (gris) ainsi que pour des simulations générées par le modèle avec les paramètres β, φ et κ estimés (noir). **c :** Critère d'information bayésien (*Bayesian Information Criteria, BIC*) calculé pour trois modèles. M1 : Softmax avec $\beta=0$, $\varphi=0$ et $\kappa>0$. M2 : Softmax avec $\beta>0$, $\varphi=0$ et $\kappa>0$. M3 : Softmax avec $\beta>0$, $\varphi>0$ et $\kappa>0$. Les valeurs de BIC plus faibles pour le modèle M3 indiquent que l'addition du paramètre φ de bonus d'incertitude génère de meilleures estimations. *Adapté de Dongelmans 2021.*

La valeur de l'option associée à la probabilité 50% est particulièrement augmentée car est associée à la plus grande incertitude. Les prédictions issues de ce modèle reflètent les choix des animaux dans le cadre probabiliste (Figure 15Ac), soutenant l'idée que l'incertitude permet de diriger l'exploration.

Dans une des publications composant la partie Résultats de cette thèse et s'intéressant à la tâche de complexité (cf. Article 2), il a été soulevé l'importance d'intégrer un paramètre moteur lié aux contraintes circulaires de la tâche, et qui permet de justifier le fait qu'en condition déterministe, les souris ont une propension à tourner en rond et effectuer peu de demi-tours, biais qu'elles conservent bien que dans une moindre mesure dans la tâche de complexité. Dans une publication plus récente de l'équipe dans la tâche probabiliste (Figure 15Bc) (cf. Annexe 1), le modèle Softmax avec β et φ a donc été agrémenté d'un troisième paramètre κ (coût lié à l'effort d'un changement de direction ou demi-tour), qui améliore encore la robustesse du modèle. Afin d'explorer les nouvelles contingences de l'environnement lors d'un changement de règle et d'améliorer leur performance, qu'il s'agisse d'une transition de la règle déterministe vers la règle probabiliste ou vers la règle de complexité, les souris accroissent leur taux de demi-tour. Elles vont donc à l'encontre de cet effort moteur (diminution du κ) pour améliorer leur représentation de l'environnement et optimiser leur comportement de choix, ce qui est un argument supplémentaire à un changement global de stratégie (ou politique) lors d'un changement de règle.

IV) Enregistrements électrophysiologiques chez l'animal éveillé dans la tâche

Afin d'étudier les bases neurales de la prise de décision et des actions auto-générées dans cette tâche, j'ai enregistré au cours de ma thèse l'activité électrophysiologique de trois structures du circuit mésocortical : l'activité multi-unitaire des neurones DAergiques de la VTA, l'activité oscillatoire de l'OFC et celle du PL PFC. Les enregistrements ont été effectués i) sur des souris mâles adultes de phénotype sauvage (WT, *wild type*) issus de la lignée C57Bl/6J (Janvier) soumis aux différents protocoles dès l'âge de 8 semaines, et ii) sur des souris DATiCre mâles adultes soumis aux différents protocoles dès l'âge de 12 semaines. Tous les enregistrements ont été réalisés à l'aide d'un système multi-canal (Neuralynx 32 canaux), à une fréquence d'acquisition de 32 kHz. Je vais ici détailler les principes des différentes méthodes d'enregistrements utilisés.

1) Activité multi-unitaire des neurones dopaminergiques

Afin de recueillir l'activité multi-unitaire (MUA, *multi unit activity*) des neurones DAergiques de la VTA, et d'être en mesure de trier les PA enregistrés, des polyélectrodes sont utilisées, associées à un système de micro-descendeur développé au laboratoire. Le signal est filtré entre 600 et 6000 Hz.

Deux octrodes sont tressées, chacune à partir de 8 fils de Nichrome (80% Nickel - 20% Chrome) torsadés. Les fils utilisés sont d'un diamètre de 17µm et enveloppés d'une gaine en Formvar amenant l'ensemble à un diamètre de 25µm. Les deux octrodes sont placées dans un microdescendeur, puis collées l'une à l'autre. Ce système permet, une fois l'implant installé de manière chronique sur l'animal, d'enregistrer des neurones à plusieurs profondeurs tout au long de son conditionnement.

L'implantation de ce dispositif se déroule sous anesthésie gazeuse à l'isoflurane (1-3% à 1L/min). Il est implanté à l'aplomb de la VTA (AP de -3,1 à -3,25 ; ML de ± 0,4 à 0,5 ; DV de -4,1 à 4,25 à partir du cerveau). Après récupération de la chirurgie, il est possible de descendre les octrodes en tournant la micro-vis se situant au-dessus du dispositif. Notons que l'activité des neurones DAergiques chez la souris éveillée diffère quelque peu de celle observée et abondamment caractérisée chez la souris anesthésiée (cf. Chapitre 2 – I) – 3)). En particulier,

les neurones DAergiques ont une fréquence et un taux de PA dans les bouffées (%SWB) plus élevés que lorsque l'animal est anesthésié.

L'identification des neurones DAergiques est tout d'abord basée sur les caractéristiques électrophysiologiques suivantes :

i) La fréquence de décharge moyenne est comprise entre 1 et 12 Hz et est très régulière.

ii) Des bouffées de plusieurs PA successifs se superposent transitoirement à cette activité de fond. Bien que la fréquence d'apparition de ces bouffées varie d'un neurone à un autre, le %SWB est quant à lui assez stable aussi pour un neurone donné.

iii) Les autocorrélogrammes ont une forme assez évasée avec une période réfractaire prononcée avoisinant les 20 ms.

Dans des enregistrements juxtacellulaires sur souris anesthésiée (Nguyen et al., 2021; Tolu et al., 2013), les neurones DAergiques présentent un PA qui peut être bi ou triphasique et d'une durée supérieure ou égale à 1,1ms. Cependant, le signal enregistré par la technique d'électrophysiologie multi-unitaire ave des multi-électrodes résulte d'un agrégat complexe d'évènements électriques de nombreuses cellules environnantes. S'il est possible de détecter des variations de potentiel extracellulaires attribuables à des PA, il est cependant malaisé d'utiliser la forme des PA comme un critère d'identification. S'il est vrai que les neurones DAergiques tendent à avoir un PA plus large, nous n'en avons pas fait un critère d'inclusion dans cette étude.

Les neurones DAergiques potentiels (pDA, *putative dopaminergic*) identifiés à l'aide de ces critères sont ensuite soumis, à l'issue d'une ou plusieurs sessions, à un test pharmacologique par administration intra-péritonéale (i.p.) de :

i) Quinpirole (1 mg/kg, 0,2 µL injectés), agoniste des récepteurs à la DA de type D2, qui diminue l'activité du neurone pDA d'au moins 30% en moins de 10 min, suivi de

ii) Eticlopride (1 mg/kg, 0.2 µL injectés), antagoniste des récepteurs de type D2, qui restaure la fréquence de décharge du neurone pDA en moins de 10 min.

Cette caractérisation se fait tout d'abord en ligne, puis est confirmée hors ligne par seuillage et *clustering* à l'aide de routines d'analyses développées par le laboratoire sur Matlab. En particulier, ces critères permettent de créer un set de neurones pDA sur lesquels toutes les analyses ont été réalisées. Les neurones ne correspondant pas à ces critères n'ont pas été utilisés dans l'analyse.

2) Activité d'ensemble des cortex préfrontaux

Afin de recueillir l'activité d'ensemble de populations neuronales dans les cortex préfrontaux, des électrodes bipolaires sont utilisées.

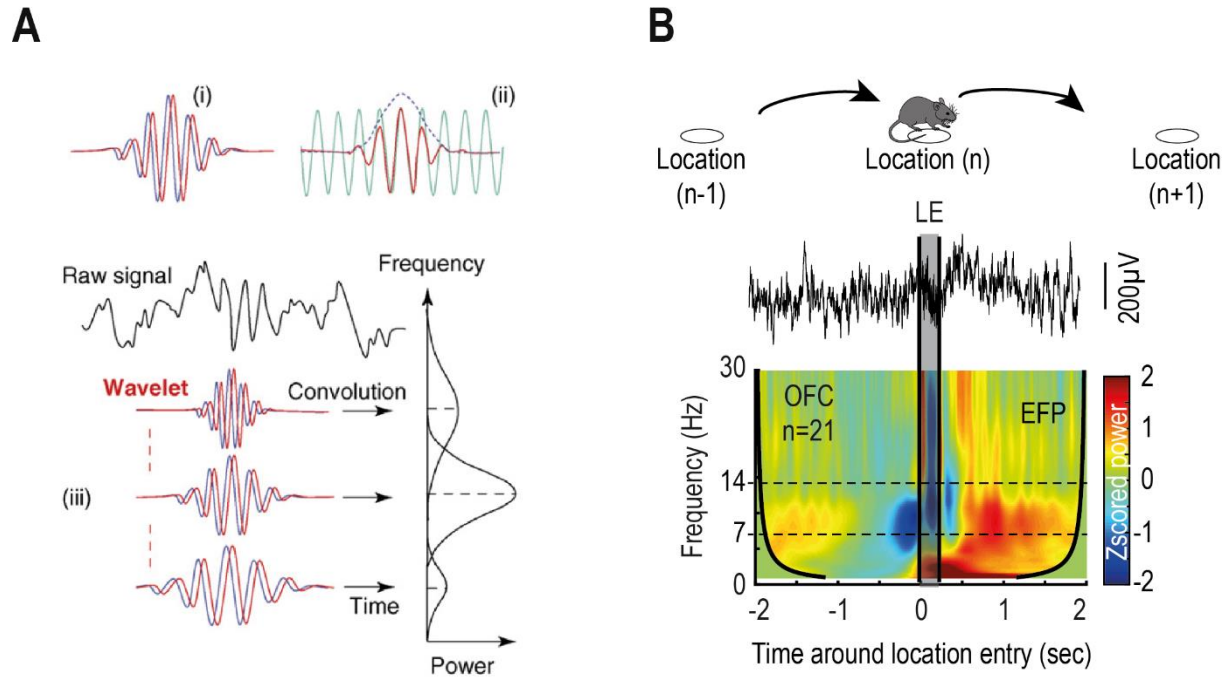
Ces électrodes bipolaires sont constituées de deux fils tressés d'acier inoxydable d'un diamètre de 50 µm et enveloppé d'une gaine de téflon amenant l'ensemble à un diamètre de 114 µm. Les deux extrémités de ces électrodes sont espacées de 0,5 mm, l'une servant de pôle de référence, l'autre de pôle d'enregistrement. Cette configuration permet par la même occasion d'éviter que les variations oscillatoires de l'EFP proviennent de rythmes résonnants depuis d'autres structures lointaines, (comme par exemple le θ locomoteur de l'HPC), puisqu'ils sont captés par les deux extrémités des électrodes et donc soustraits. Elles permettent ainsi d'encadrer une colonne corticale et d'obtenir un signal reflétant l'activité locale.

L'implantation de ces électrodes se déroule sous anesthésie gazeuse à l'isoflurane (1-3% à 1L/min). Elles sont implantées à l'aplomb du lOFC (AP +2.6 mm, ML ±1.5 mm, DV -1.7 mm) et du PL mPFC (AP +1.65 mm, ML ±0.5 mm, DV -1.8 mm), puis fixées définitivement à la surface du crâne.

Le signal issu de ces électrodes est double :

- Il est d'une part filtré de 0,1 à 300 Hz afin de recueillir les variations oscillatoires de l'EFP. L'analyse spectrale du signal oscillatoire a été faite à l'aide d'une transformée en ondelettes de Morlet (Le Van Quyen and Bragin, 2007). Ce type de transformation de signal permet en effet d'avoir accès aux dynamiques temporelles fines du signal oscillatoire, et est parfaitement adapté dans la tâche 3 points dans laquelle les comportements observés sont d'ordre inférieur à la seconde. Une analyse de Fourier s'applique plutôt à des signaux supposés stationnaires, et même fenêtrée (*short-term Fourier transform*), elle ne permettrait pas une telle précision temporelle. La transformée en ondelettes est un produit de convolution entre le signal EFP et des formes plus ou moins dilatées d'ondelettes, qui sont, dans le cas des ondelettes de Morlet, le produit entre une sinusoïde et une gaussienne. Une ondelette aura donc un produit de convolution maximal avec une oscillation transitoire. Effectuer la transformée avec des versions compressées (ou étirées) d'ondelettes permet de détecter des oscillations de périodes plus ou moins courtes.

Cela permet donc d'étudier respectivement les composantes de fréquences hautes ou basses du signal.



<u>FIGURE 16 :</u> Analyse spectrale par transformée en ondelettes de Morlet

A. Les ondelettes sont des oscillations localement périodiques. L'ondelette mère (i) est une ondelette de Morlet qui est une one sinusoïdale complexe multipliée par une enveloppe Gaussienne (ii). Par compression ou étirement de l'ondelette mère (i) peut être obtenue une famille d'ondelettes de telle sorte qu'elles échantillonnent de manière adéquate toutes les fréquences présentes dans le signal. Une ondelette « dilatée » permettra l'analyse des oscillations de basses fréquences et les ondelettes « contractées » des hautes fréquences. La convolution de ces ondelettes sur le signal reflète leur recouvrement, et est appliquée pour chaque ondelette, c'est-à-dire pour chaque fréquence (iii). *Adapté de Le Van Quyen & Bragin 2007.*

B. Exemple d'analyse des EFP dans l'OFC par transformée en ondelettes de Morlet dans la tâche 3 points, à la fin du cadre déterministe. Cette transformation du signal EFP de l'OFC permet de représenter les variations de puissance par fréquence (ici entre 0 et 30 Hz) au cours du temps (ici t=0 sec correspond à la détection de l'animal sur une cible, la bande grise correspondant aux 200ms que dure l'ICSS). Dans cet exemple, les essais sont moyennés par session et sur l'ensemble des animaux. Le spectre de puissance est normalisé sur la période de 2 secondes précédant la détection de l'animal sur une cible, et ses variations de puissance sont représentés en code couleur (*Zscored power*). Cette analyse révèle une augmentation de puissance dans la gamme θ (7-14 Hz) au moment où l'animal se dirige vers le prochain point (cf. Article 1). Notons qu'il est également possible de filtrer le signal dans la gamme de fréquence d'intérêt (7-14 Hz) afin de représenter les variations de puissance au cours du temps en deux dimensions.

- Le signal est d'autre part filtré entre 600 et 6000 Hz afin de recueillir les PA des populations neuronales de l'OFC et du PFC. Puisqu'issue d'un unique point d'enregistrement, cette activité de population ne permet pas de séparation d'unités neuronales (*clustering*) (comme c'est le cas de la MUA enregistrée dans la VTA). Les PA sont extraits par seuillage et l'activité en fréquence moyenne est calculée, reflétant donc une activité d'ensemble de la population neuronale dans la région implantée.

Chapitre 5 : Première publication

Self-generation of goal-directed choices in a distributed dopaminergic and prefrontal circuit. Bousseyrol E[1,2#], Didienne S[1,2#], Takillah S[1,2], Prévost-Solié C[1,2], Come M[1,2], Ahmed Yahia T[1,2], Tricoire L[2], Dalkara D[3], Mourot A[1,2], Naudé J[1,2†*] & Faure Ph.[1,2†*]. In prep.

Nombre des comportements déployés par les animaux sont dirigés vers des buts et motivés par leurs conséquences notamment en termes de récompense. Les comportements orientés peuvent être déployés en réponse à des stimuli externes de l'environnement, mais peuvent également découler de représentations dynamiques et internes à l'animal. L'étude des mécanismes neuronaux de la prise de décision a largement reposé sur l'utilisation de stimuli sensoriels afin de spécifier à l'animal les différentes options disponibles ou le moment auquel agir. Cela a conduit à considérer les circuits neuronaux impliqués comme un ensemble de modules effectuant de façon sérielle (du stimulus à l'action) des calculs indépendants.

Le travail de cette publication a eu pour objectif de questionner le rôle du circuit mésocortical dans les comportements de choix auto-générés chez la souris, en particulier dans un environnement probabiliste. Il a été réalisé dans une tâche à choix multiples dans laquelle les animaux effectuent des séquences de choix de manière auto-générée (cf. Chapitre 4). Cette tâche nous a permis de quantifier de manière fine à la fois les choix des animaux basés sur les valeurs subjectives des différentes options disponibles, mais également la motivation liée à ces choix.

Dans cette étude, nous avons enregistré l'activité électrophysiologique des neurones dopaminergiques (DA) de l'aire tegmentale ventrale (VTA) ainsi que l'activité de populations neuronales dans le cortex orbitofrontal (OFC) et préfrontal (PFC) chez l'animal vigile, afin d'étudier les dynamiques d'ensemble du circuit mésocortical dans les décisions des animaux. A l'aide de perturbations spécifiques de ces structures, nous avons identifié une séquence d'oscillations et d'activité neuronale dans la VTA, l'OFC et le PFC qui co-encode et co-détermine l'auto-initiation et les choix. Cette séquence d'activités apparait avec l'apprentissage de la tâche par une réorganisation dynamique d'activités neuronales spontanées, et dépend du contexte et de l'incertitude associée aux différentes options. Nos résultats montrent que l'OFC

et la VTA interagissent pour initier des actions dirigées. Par ailleurs, l'utilisation d'un modèle de prise de décision dérivé du Softmax et adapté à la structure de notre tâche (cf. Chapitre 4) nous permet de conclure que le PFC et la VTA participent à la motivation des décisions auto-générées ainsi qu'à la politique de choix déployée par les animaux face à l'incertitude.

Self-generation of goal-directed choices in a distributed dopaminergic and prefrontal circuit

Bousseyrol E[1,2#], Didienne S[1,2#], Takillah S[1,2], Prévost-Solié C[1,2], Come M[1,2], Ahmed Yahia T[1,2], Tricoire L[2], Dalkara D[3], Mourot A[1,2], Naudé J[1,2†*] & Faure Ph[1,2†*]

Affiliations

[1] Sorbonne Université, INSERM, CNRS, Neuroscience Paris Seine - Institut de Biologie Paris Seine (NPS - IBPS), 75005 Paris, France.

[2] Brain Plasticity Unit, CNRS, ESPCI Paris, PSL Research University, 75005 Paris, France

[3] Sorbonne Université, INSERM, CNRS, Institut de la Vision, Paris, France.

[#] These authors contributed equally

[*] These authors contributed equally:

[†]Correspondence to phfaure@gmail.com and jeremie.naude@gmail.com

Summary

Goal-directed choices that are not triggered by external cues arise from internal representations of the outcomes. The use of a stimulus to specify when to act, which option to take, or whether to explore, has led to consider the reward circuit as a feedforward set of modules carrying independent computations. Here, we develop a task in which mice self-determine the initiation, direction, vigor and pace of actions based on their memories of the outcomes. Using recordings and perturbations, we identify a sequence of oscillations and firing in the ventral tegmental area (VTA), orbitofrontal (OFC) and prefrontal cortex (PFC) that co-encodes and co-determines self-initiation and choices. This sequence appeared with learning as an unguided realignment of spontaneous dynamics, and the interactions between the structures depended on the reward context, and in particular the uncertainty associated with the different options. We suggest that self-generated choices arise from a distributed circuit based on an OFC-VTA core setting whether to wait or to initiate actions, with the PFC engaged by reward uncertainty to participate in both the selection and pace of actions. (~175 words)

Highlights

- Self-paced actions arise from contextual reorganization of mesocortical dynamics.
- VTA, PFC and OFC complementarily encode predictions and errors about outcomes.
- Distributed firing-then-oscillations dynamics set the goal, initiation and pace of actions.
- VTA and PFC antagonistically promote and inhibit motivation by reward uncertainty.

Introduction

A major distinction is made between respondent and operant behavior (Passingham 2010, Costa 2011). Actions can be triggered by an external stimulus and the behavior is said to be respondent. But often, animals base their decisions solely on internal representations of their goals. In such case, the behavior is said to be operant but also "self-paced", “self-generated” or “self-initiated” (Klaus et al., 2016; Passingham et al., 2010). This includes decisions like when to act, but also like which option to take in cases where no stimulus indicates each option’s value. Decision-making in the absence of cues thus necessitates to compare internal memories of the values associated with different options. Finally, when faced with several alternatives, animals do not always exploit the option associated with the highest expectation of reward, but instead explore less rewarded options, in order to gain information (Cohen et al., 2007). Again, while exploration might be guided by novel or salient cues, it can also be a consequence of an internal exploratory mode (Cohen et al., 2007). However, most of our knowledge on decision-making has been obtained with respondent, stimulus-triggered behaviors that are most reproducible and amenable to statistical analyses (Passingham 2010, Costa 2011). Hence, the neural mechanisms by which animals initiate goal-directed actions, decide between them, and explore potentially informative ones in the absence of external cues, is far from being understood.

The choice of the experimental paradigm through which the neural mechanisms of goal-directed choice are probed is not neutral. In respondent behavior, the stimulus constitutes an identified input to the nervous system and the action its output (Gomez-Marin 2019). In this framework, it is tempting to treat the meso-cortico-limbic decision circuit as a feedforward process. Hence, a long-standing question has consisted in assigning a particular and sequential role to each stage of the circuit (Hunt and Hayden, 2017), e.g. to the ventral tegmental area (VTA) midbrain dopamine nucleus, the frontal areas (i.e. prefrontal and orbitofrontal cortex) and basal ganglia. This is illustrated by the reinforcement-learning computational theory, which identifies phasic release of dopamine (DA) with a reward prediction error (RPE), i.e. the comparison between actual and expected reward. DA RPE would constitute a teaching signal for learning the appropriate stimulus-action responses (Schultz, 2007; Schultz et al., 1997). In this theory, DA would only play an indirect role in behavior, affecting subsequent trials of a task rather than the current one. In the modular view, cortices provide subcortical areas with information about the current state and options of the environment and the basal ganglia then select among these options and initiate the goal-directed action (Gershman and Uchida, 2019; Klaus et al., 2016). Hence, similar to DA, frontal cortices would only indirectly affect decisions, by signaling sensory states, beliefs and possibilities to the basal ganglia, and/or by conveying value predictions to be compared with actual outcomes by DA neurons (Starkweather et al., 2018; Takahashi et al., 2011).

However, both dopamine and frontal areas have been described as having more direct roles in decision-making, and the basal ganglia might not be the only locus of action selection (Cisek, 2007). Good-based models place the choice process at the level of the OFC/PFC (Padoa-Schioppa, 2011), as frontal cortices not only represent the

value of potential options but also actively compare between them (Cai and Padoa-Schioppa, 2014; Hunt and Hayden, 2017). It is however unclear how economic choice is independent from action planning (Cai and Padoa-Schioppa, 2014) or result from recurrent computations incorporating motor parameters of action execution back into a recurrent decision process (Cisek, 2012). Theories on cognitive control assign a top-down, potentially inhibitory role to the OFC and PFC through computing goals, plans and task rules, i.e. higher-order decision-making (Miller and Wallis, 2013). These different accounts all point at a direct, active role of frontal areas on ongoing choices. Phasic DA also exerts an impact on the ongoing behavior and on self-paced actions, by modulating the vigor of actions leading to reward (Berke, 2018; Salamone and Correa, 2012). Mixed results have been obtained on DA facilitating action self-initiation itself, depending on the type of task, DA nuclei, and intensity of DA manipulation (Coddington and Dudman, 2018; Fischbach-Weiss et al., 2018; Howe and Dombeck, 2016; Silva et al., 2018). Given the respective roles of DA and frontal cortex in value-based decision making, selection and action initiation, the question of the coordination between these circuits arises, in learning processes, in decision-making and exploration, and in motor execution.

Overall, studies focusing on prefrontal modulation of the VTA have focused on learning (Jo and Mizumori, 2015; Lak et al., 2020; Starkweather et al., 2018; Takahashi et al., 2011) while work assessing dopamine modulation of frontal cortices have pointed at decision-making (Ellwood et al., 2017; Puig et al., 2014; Verharen et al., 2018). As the mesocortical circuit is a loop rather than a feedforward process (Haber and Knutson, 2010; Hunt and Hayden, 2017), an integrated view is thus lacking. Concurrent recordings of VTA and frontal cortices (Fujisawa and Buzsáki, 2011; Park and Moghaddam, 2017) have shown synchrony between these structures but have not delved into causal mechanisms. More importantly, these studies have focused on the role of DA or cortical dynamics in cue-guided behaviors (Fujisawa and Buzsáki, 2011; Jo and Mizumori, 2015; Lak et al., 2020; Park and Moghaddam, 2017; Puig et al., 2014; Starkweather et al., 2018; Takahashi et al., 2011), or, when assessing goal-directed actions, did not assess the respective roles of these structures in the self-initiation (Ellwood et al., 2017; Verharen et al., 2018). In the present study, we used an experimental paradigm in which mice must perform a sequence of choices to obtain rewards associated with intracranial self-stimulation (Belkaid et al., 2020; Dongelmans et al., 2021; Naudé et al., 2016). This protocol satisfies a series of requirements that have proven difficult to address altogether: simultaneously monitoring action selection, action execution, and electrophysiological activity of the mesocortical circuit, in an environment with controlled rewards and exploration. Most importantly, mice generate a vast number of “template” bouts of locomotion in this task (Naudé et al., 2016), at a constant motivational level insured by intracranial self-stimulation, thus alleviating the need for a stimulus to repeat the behaviour.

In this task, we show that goal-directed actions from one rewarded location to the next consisted in stereotyped, ballistic (bell-shaped speed profiles) movements that nonetheless presented characteristic variabilities in their timings or directions. Mouse behavior displayed hallmarks of self-generated actions, with initiation, pace and decisions underdetermined by environmental cues, but depending on internal representations of the reward

context. We further show that the VTA, PFC and OFC coordinated their activities into a sequence of distributed firing and oscillations, associated with the goal-directed actions. Such sequence emerged with learning as a reorganization of existing dynamics to lock at behaviorally-relevant timings, without being triggered by a stimulus. Combining electrophysiology with optogenetic and pharmacological manipulations, we unveil that OFC-VTA interactions set self-initiation in each reward context, with a more specific involvement of the PFC during decisions under uncertainty and exploration. We also suggest that the VTA and PFC can act in synergy to self-pace the actions, but may have antagonistic roles in pondering the influence of uncertainty on choices, in particular for exploration. Our study highlights how the mesocortical circuit self-generates decisions through distributed yet distinct computations, by reorganizing itself in a context-dependent manner.

Results

Mouse actions underdetermined by environmental cues, but shaped by internal representations, indicate self-paced decisions.

To decipher the neural bases of self-paced decisions, we used a spatial version of a multi-armed bandit task adapted to mice (Belkaid et al., 2020; Dongelmans et al., 2021; Naudé et al., 2016). Three equidistant locations explicitly marked in an open-field were associated with rewards delivered as intra-cranial self-stimulations (ICSS) in the medial forebrain bundle (MFB) (Carlezon and Chartoff, 2007) (Figure 1A). MFB ICSS eliminates the need for food restriction and the associated satiation level, which is known to affect decisions, in particular under reward uncertainty (Schuck-Paim et al., 2004). Mice could not receive two consecutive ICSS at the same location, they therefore alternated between rewarding locations. Their behaviour in the task thus consisted in sequences of trials in-between rewarding locations (Figure 1A). First, on each rewarded location, mice could either circle forward along the three locations, or perform U-turns by coming backwards, at trial i+1, to the i-1 location (Figure 1B). Second, a trial is characterised by a "template" bout of locomotion: an initiation of a movement towards the next target, an acceleration followed by a slowing down and a pause at the next target (Figure 1C left). Execution of this ballistic (bell-shaped) velocity profile was characterized by an important trial-to-trial variability. First, the successive timings of the behavior presented substantial variability: the time during which animals dwelled at a rewarded location before initiating a new trial, the time to reach maximal speed, and the overall time to go from one location to another (Figure 1C right). The direction (U-turn or Forward) impacted on the dwell duration (which is an early trial timing), and finally on the time to goal (Figure 1D). It indicated that early decisions impacted on the later ones and that the successive timings were nested within the behavioral sequence. More precisely, multiple linear regression (see Methods) showed that the trajectory direction, the dwell time, and the time to maximal speed all had significant impact on the time to goal (Figure 1E). This successive addition of variability at each stage of a trial suggests that trajectory direction, initiation, and vigor were underdetermined by external cues (Klaus et al., 2016; Passingham et al., 2010). We thus assessed how these decisions depended on animals' internal representations of the potential rewards by comparing behaviors in two reward contexts. Mice were initially trained in a deterministic context (D) in which all locations were associated with a certain ICSS delivery (P = 100%). Then, mice underwent a probabilistic (P) context, in which each target was associated with a different probability of ICSS delivery (here; P = 100%, 50%

and 25%; Figure 1E). The time to goal (Δ=1.96, p=0.001) and dwell time (Δ=0.26, p=0.02) increased in the P setting compared to the D setting, suggesting a decreased motivation due to less frequent reward delivery in the P setting (Figure 1F). U-turn behavior (Δ= 0.19, p<0.001) increased as well, reflecting an adjustment of the trajectory directions to the respective payoffs of the locations: while mice visited the three targets uniformly in the D setting, in the P setting they visited more often the targets associated with the highest ICSS probabilities (i.e. p_{100} and p_{50}, Figure 1G). Hence, mice behavior displayed hallmarks of self-paced decisions: direction, initiation and vigor of actions were underdetermined by external cues, but instead influenced by internal representations of the actions' potential outcomes.

Distributed mesocortical sequence associated with self-generated decisions depends on reward context

As self-generated behaviors are caused by changes in the internal state of the animals, rather than by external changes in the environment (Klaus et al., 2016; Passingham et al., 2010), we next characterized the mesocortical dynamics associated with self-paced decisions. We recorded from putative dopamine neurons (pDAn, n=136 neurons) in the VTA from WT mice (n= 12 mice) using extracellular multi-electrodes (Figure 2A Left). All neurons met the electrophysiological and pharmacological criteria used to identify dopamine cells in vivo (Roesch et al., 2007; Takahashi et al., 2017) (Methods). In another group of mice, bipolar electrodes were also chronically implanted bilaterally into the PFC and OFC (n= 21 mice), allowing to record both extracellular field potentials (EFP) and population spiking activity (Figure 2A right, Methods).

At the end of the D setting, pDAn activity (n=65 cells) emitted bursts of action potentials early in the trials (henceforth called "early bursting"), right after the ICSS delivery, during the dwelling period (Figure 2B). Early bursting in VTA pDAn was paralleled by an early increase in OFC population firing (Figure 2C). After the dwell time, as the animals started moving towards the next target (0.5-1 sec window), oscillation power in the θ band (7-14 Hz) increased in the OFC EFP (t-test p=0.00013, Figure 2C). In contrast, no specific activity was observed in the PFC, neither in the population spiking nor in the EFP (Figure 2D). Consistent with this temporal order, OFC firing (whether considering all the spikes or just the increased activity) did not display phase-locking to θ oscillations in the OFC, but the two signals displayed a lagged cross-correlation (Figure 2E), indicating a transition from increased spiking to θ oscillations around the self-initiation of the trial by the mice. We next asked how mesocortical activity changed with the reorganization of animals' choices when faced with reward omission in the probabilistic context. While pDAn early bursting was still present at the end of the P setting, the increase in firing frequency was lower than what was observed in the D setting. The parallel increase in OFC firing during dwelling, and latter power increase in θ oscillations, were similar in both settings. By contrast, an activity emerged in the PFC specifically in the P setting, consisting in an increase in both PFC population firing and δ (3-6 Hz) oscillation power (0.4-0.9 sec window) (Figure 2H). Overall, the mesocortical network, classically associated with stimulus-triggered decision-making (Jo and Mizumori, 2015; Lak et al., 2020; Starkweather et al., 2018), was also recruited during self-generated decisions, but with a specific involvement of the PFC during decisions under uncertainty. Our results are consistent with the view that self-generated decisions do not need a separate neural substrate but might reuse the neuronal circuits and computations of cue-guided behavior, yet in an 'internally generated' mode of brain dynamics (Pezzulo and

Cisek, 2016), with the PFC exerting context-dependent control of the strategy when choice values are uncertain (Sharpe et al., 2018).

Self-generated mesocortical sequence emerges with learning as a reorganization of existing dynamics

The dependence of VTA, PFC and OFC activities on the reward context suggests that the mesocortical dynamics associated with self-paced actions reorganizes based on the action's outcomes. We thus investigated how this dynamic first emerged with learning in naïve mice within the D context. Along the sessions of the D context, the total number of reward locations visited per session increased (Figure 3A), confirming place reinforcement. The averaged time to goal ($p<0.001$, Figure 3B) and the dwell time ($p<0.001$, Figure 3B) decreased accordingly, while the maximal speed increased ($p<0.001$, Figure 3B). Finally, the proportion of U-turn decreased (Figure 3C, $p<0.001$), indicating that mice learned to optimize their trajectories, and reduce the motor cost associated with U-turns (Belkaid et al., 2020).

This behavioral learning was associated with a modification of VTA pDAn dynamics. The proportion of pDAn, with a firing rate significantly higher than baseline (see Methods) in-between two rewarded locations, increased with the training sessions (Figure 3C). Early in learning, when the behavior is still dominated by spontaneous locomotion, fluctuations in pDan firing rates were not locked to any task-related events (thus averaging out). In contrast, at the end of learning, increased pDAn activity appeared at similar timings across trials, as reflected in the significant increase when averaged (Figure 3C). However, the correlated locking to behavioral timings in dopamine firing during a trial, and increase in the number of trials with learning, did not result in an overall (session wide) increase in pDAn firing frequency ($F_{(2,13)}=1.3$, $p=0.28$, Figure 3D), but rather in an increase of the pDAn bursting pattern (Figure 3D, %SWB: $F_{(2,13)}=$, $p=0.04$). Hence early bursting in dopamine neurons did not rely on additional spikes, but rather on a dynamical re-organization towards bursting activity at behaviorally relevant times. Similarly, the total number of OFC EFP in which θ oscillation power was higher than baseline in-between two rewarded locations gradually increased with learning (Figure 3E). Oscillatory activity in θ band emerged around the same timing across trials, from relatively scarce task-locked θ oscillations in early learning (Figure 3E). When compared with power spectra from naïve animals, θ power did not significantly change throughout the learning of the task. Rather than occurring more often, θ oscillations reorganized to lock at particular timings of the behavioral sequence. By contrast, no change in PFC activity was observed (Fig Supp), which is expected as PFC does not display δ oscillations at the end of the D setting (Figure 2). Overall, during the learning in the deterministic context, the optimization of trajectories and speed profiles were associated with a reorganization of VTA DA cells towards time-locked bursting and of OFC towards time-locked θ oscillations.

Bursting of VTA DA cells, together with frontal firing and oscillations, form a distributed signal for outcome discrepancy and expectations.

Learning theories propose that behavior, and underlying brain activity, reorganize when unexpected outcomes occur: the discrepancy between observed and expected reward may be used to update the animals' internal representations (Dayan and Daw, 2008; Rescorla and Wagner, 1972; Sutton and Barto, 1998). This comparison is

thought to be computed in mesocorticolimbic areas (Takahashi et al., 2009, 2011; Yun et al., 2020) , most notably in VTA DA cells (Schultz et al., 1997). To assess the reward representation generated by the intra-cranial stimulation, and disentangle its direct effect on mesocortical activity, either as an artifact or as a reward response, we thus evaluated VTA, OFC and PFC activity during expected rewards, and unexpected rewards and omissions. As previously described for the end of the D and P contexts (Figure 2), early bursting activity in DA cells at the beginning of the P context occurred during the dwelling period and before the action initiation (Figure 4 A top row). This increased activity was observed even when the ICSS reward was unexpectedly omitted at the beginning of the P setting (Figure 4 B top row), but did not appear after unexpected, random stimulation reward in the home-cage (Figure 4 C top row). This suggests that, rather than being generated by the previous reward, early bursting at the beginning of the trial may be related to the expectation of the upcoming reward (Sutton and Barto, 1998; Wassum et al., 2012), or the invigoration of the next movement (Howe and Dombeck, 2016; Silva et al., 2018). Moreover, pDAn activity decreased at the time of the expected reward during unexpected omissions (Figure 4 B top row) and increased during unexpected stimulation reward (with more reward than expected, Figure 4 C top row), consistent with pDAn computing a reward prediction error (Glimcher, 2011; Schultz, 2007; Sutton and Barto, 1998). Besides, the decreased pDAn activity following random stimulation in the home-cage emphasizes the emergence of early bursting activity in pDA cells specifically upon self-paced actions. Likewise, in the OFC, neither the increased power of θ oscillations (Figure 4A second row) nor the increased firing (Figure 4A third row) at the beginning of the trials were triggered by the previous stimulation reward, as both θ oscillations and increased firing were observed during omission trials (Figure 4B). Furthermore, unexpected, random stimulation rewards did not generate θ oscillations (Figure 4C), indicating a specific involvement in self-paced behavior. Unexpected reward increased OFC population firing, suggesting an influence of both expected and unexpected outcomes on OFC firing activity. Finally, in the PFC (Figure 4 fourth and last row), the δ oscillations and increased firing observed in the P setting (Figure 2), but not in the D setting (Figure 4A) could already be observed in the first omission trials (1-1.5sec post omission, Figure 4B fourth and last rows). By contrast, unexpected stimulation reward decreased δ oscillations and increased population firing, suggesting an involvement of the PFC specifically following unexpected outcomes.

Distributed and complementary representations of decision parameters in mesocortical structures.

The mesolimbic dynamics emerging with decisions were not only caused by the immediately preceding outcome. We thus searched for a signature of the internal representations of the upcoming outcome, which is thought to guide self-paced decisions (Klaus et al., 2016; Passingham et al., 2010). First, we determined which of the task parameters actually affect the animals' choices. To do so, we formalized how mouse decisions adapted to the different expected payoffs of the locations in the P setting, and to the motor requirements of U-turns.

As mice could not receive two consecutive rewards at the same location, they faced a succession of binary choices between the two remaining locations (Figure 5A left). We thus expressed choices in the P setting as the proportion of exploitative choices in three gambles (G) between the potential outcomes (Figure 5A, G_{25}:100 vs 50, G_{50}: 100 vs 25 and G_{100} 50 vs 25% reward probabilities, respectively, exploitative choices consisting in choosing

the option with the highest reward probability). Mice displayed a preference for higher reward probabilities in G_{100} ($p=0.02$) and G_{50} ($p=3.10^{-4}$), but not in G_{25}, in which they chose the locations associated with 50% and 100% reward in the same proportions ($p=0.20$). This replicated our previous studies (Dongelmans et al., 2021; Naudé et al., 2016): in the P setting, mice assigned a positive value to uncertainty, which is zero for predictable outcomes (100% probability) and maximal for unpredictable outcomes (50% probability). As reward expectation (probability) and uncertainty (variance) co-vary in this setup, we used a model-based analysis to disentangle the influence of expected reward and uncertainty on choices. We compared increasingly complex models of decision-making to assess which one best explained the experimental data (see Methods). We also included a motor cost as a potential parameter affecting trajectory directions, as mice favored forward trajectories rather than U-turns (Figure 1), in both the D and P settings. Model comparison confirmed that in the P setting, animals added the uncertainty to the expected reward as a total positive value, discounted by the motor cost (negative value) of U-turns (Figure 5B).

We thus assessed the encoding of expected reward, uncertainty and motor cost in mesocortical activity to understand how outcome representations shaped self-direction. To do so, we split trials according to either the trajectory (forward or U-turns), the expected reward, and uncertainty, of the target locations. We did not find any evidence for motor cost in any of the recordings (Figure 5C). By contrast, VTA pDAn activity scaled with expectation of reward (Figure 5D, $p =4.10^{-5}$, one-way ANOVA). This finding is consistent with early bursting in pDAn signaling a "time-difference" reward prediction error (Sutton & Barto): in this framework, DA cells not only signal the difference between actual and expected reward, but also integrate the expectation of future rewards predicted by the current state and actions of the animal (Schultz et al., 1997). Early bursting scaled with the reward probability of the chosen outcome, consistent with a quantitative time-difference reward prediction error (Fiorillo et al., 2003).

PFC δ oscillations power did not scale with the expectation of reward uncertainty (Figure 5E), in contrast to PFC population firing (Figure 5F) that was maximal when the mice directed towards the location associated with the most uncertain (50%) reward probability. Hence, not only PFC δ oscillations reflected previous unexpected outcomes (Figure 4), but also population firing signaled the expected uncertainty associated with the predicted outcome. Finally, we did not find evidence for OFC θ oscillations (nor firing) to encode for expected reward or uncertainty. OFC oscillations were uncorrelated to the expectation of outcome value. They occurred following both reward and omission during the task, but not after passive exposition to unexpected rewards in the home-cage (Figure 4). All of this suggests a specificity related to active behaviors, rather than to the expectation of outcomes. Hence, internal representations of reward outcomes that influenced animals' self-directions, i.e. expected reward and uncertainty, were represented in a complementary way by VTA and PFC activities, respectively, with the OFC encoding a more general, active change in the behavioral state of the animals.

Distributed correlates of self-initiation, invigoration and pace of goal-directed actions

After examining the internal representations directing self-paced behavior, we then assessed the relation between circuit activity and the content of self-paced actions. We analyzed how the initiation and invigoration of goal-directed behaviors related to mesocortical dynamics at each trial. As stated above (see Figure 1) each trial is characterised

by the initiation of a movement towards the next target (estimated by the dwell time) and a strong acceleration (estimated by the time to maximal speed) followed by a deceleration at the next target, resulting in the overall time to goal (see Figure 1, Figure 6A top) and then another pause. Sorting, for each neuron, the trials by ascending pDAn early bursting revealed a negative correlation with the time to goal (Figure 5A) and the time to reach the maximal speed (Figure 6B-C): greater pDan activity before the self-initiation correlated with shorter time to goal, due to a shorter time to reach the maximal speed. Distribution of correlation coefficients (R^2) for all neurons suggests that among the previous decisions that influences the time to goal (initiation and speed), correlation was the strongest with the time to reach the maximal speed (Figure 6D). We ruled out that these correlations might have been spurious using surrogate data preserving the coding capacity of pDAn cells and the timing variability of mice (Figure 6D, surrogate and see Methods).

By contrast, we did not find any significant correlation between the amplitude of cortical oscillations and the successive timings of mice behaviors, neither for OFC θ nor for PFC δ (Figure 6E). This might be due to the temporal order of neural oscillations and behavioral events: the increase in OFC θ and in PFC δ generally occurred after self-initiation (see Figures 2 and 4), and thus may not be involved in this decision process. On the opposite, increase in OFC population activity, which occurred early in the trial, correlated positively with the dwell time, suggesting an involvement in the self-initiation of the trial (Figure 6F). Finally, PFC population firing correlated with the time to goal, but neither with the dwell time nor with the time to maximal speed (Figure 6G). This indicate that additional variability in the overall pace of the trial, that was not already due to earlier decisions (Figure 1), may be encoded in the PFC. Overall, OFC, VTA and PFC firing activity respectively encoded for self-initiation, invigoration, and pace of the goal-directed actions, suggesting a sequential and distributed mechanism for self-paced decisions.

Synergy and antagonism between mesocortical structures in self-paced decisions under uncertainty

Electrophysiological analyses suggest an involvement of VTA and PFC in both action selection (Figure 5) and execution (Figure 6). We thus investigated their causal involvement, by inactivating these structures in the task. To specifically manipulate VTA DA neurons, we expressed an inhibitory halorhodopsin variant (Jaws (Chuong et al., 2014)) in DAT^{iCRE} mice using a Cre-dependent viral strategy (Figure 7A). We confirmed expression of the opsin with immunochemistry (Supplementary Figure 1A) and in patch-clamp recordings, with 500ms-light pulses at 0.5 Hz (520 nm) reliably decreased their activity in Jaws-transduced mice. As the amplitude of the early bursting in VTA DA cells correlated with movement invigoration (Figure 6), we specifically tested the effect of optogenetic inactivation of VTA DA neurons at the time of the early bursting activity (i.e. 500 ms continuous, starting 100 ms after the last ICSS, see Methods, Figure 7A). Optogenetic inhibition on each trial in the D context decreased the total number of transitions (Δ= -13.8, p=0.025, Figure 7B). The decrease in the number of trials was due to an alteration in speed profile that started during light application and lasted after its termination (Figure 7C), with an increase in the overall time to goal (Δ= 2.35s, p=0.027, Figure 7D left). In particular, photo-inhibition of VTA DA cells delayed the action initiation (increased dwell time, Δ= 0.17s, p=0.020, Figure 7D middle) and decreased the action vigor (decreased maximum of mean speed, Δ= 4.81 $cm.s^{-1}$, p=0.016, Figure 7D right). None of these

parameters were modified in GFP-transduced DAT^{iCRE} mice (Supplementary Figure 1B). These results implicate VTA DA cells in the invigoration, but also in the initiation, of self-generated behaviors.

Furthermore, VTA DA photo-inhibition affected OFC θ oscillations (Figure 7E left), with a decrease in θ oscillations power during the light application (Δ= -0.09, p=0.038, Figure 7E right) which did not last afterwards. By contrast, we did not find any effect of VTA DA photo-inhibition on other frequencies of the OFC power spectra, nor in PFC oscillation, suggesting a specific relation between early DA bursting and OFC θ oscillations. As the OFC dynamics transitioned from increased population firing to θ oscillations at action initiation (Figure 2C, Figure 7E), VTA DA photo-inhibition may have delayed θ oscillations directly, or indirectly, with OFC θ oscillations merely following the delayed action initiation. This latter interpretation is consistent with the peak in OFC θ power occurring at 0.62s (and a dwell time of 0.41s) without light; while the peak in OFC θ power under VTA DA photo-inhibition occurred at 0.83s (with a dwell time of 0.59s).

The absence of observable effects of VTA DA photo-inhibition on PFC dynamics is also in line with the electrophysiological results suggesting that the PFC is not implicated in self-initiation in the D context. This prompted us to probe the involvement of the PFC in the P context, in which PFC δ has been specifically detected, and PFC firing frequency has been correlated with time to goal and reward uncertainty. We thus inactivated the PFC using a local infusion of muscimol (Figure 7F). PFC inactivation in the P context increased the time to goal (Δ=+0.951, p=0.032, Figure 7G left), which resulted in a lower number of transitions (Δ=-12.5384, p=0.02, Supplementary Figure 1C) without affecting intermediate timings of the self-paced behavior (dwell time and time to maximal acceleration). This confirmed the electrophysiological results that showed an encoding of time-to-goal, but not of intermediate timings, by PFC population firing (Figure 5), suggesting a role for the PFC in setting the overall pace of the action.

Surprisingly, muscimol also decreased the percentage of U-turns (Δ=-6.6029, p=0.0026, Figure 7G right). This was due to an alteration of choices: PFC inactivation affected choice repartition on the three locations (Δ=-1.3335, p=0.0024, Figure 7H left) with a decrease probability to visit location associated with the highest probability of reward (i.e 100%). We thus sought to explain these changes in U-turns and in the choices of the best locations using the computational model. We fitted the transition function of each mouse (Supplementary Figure 1D) in the saline and muscimol sessions with a model taking into account the value sensitivity (β), the weight of uncertainty (φ) and the motor cost (κ) of each decision (see Methods and Figure 5B). We found that under muscimol, animals behaved as if their uncertainty bonus was increased (Δ=+0.7310, p=0.0215, Figure 7H right), with no changes in β and in κ (Supplementary Figure 1E and F). Again, PFC inactivation by local muscimol extends the electrophysiogical results: PFC population firing positively encoded reward uncertainty, but PFC inactivation increased the valuation of uncertainty, suggesting an inhibitory control of PFC on uncertainty-seeking.

Discussion

By using a task in which mice perform stereotyped trials from a motor point of view (going from one rewarded location to another), we eliminated the requirement of a stimulus to reiterate and time-lock a specific behavior, necessary for statistical analyses. The absence of a cue specifying the direction or the initiation of a trial however generated sufficient variability (mice had to choose among several alternatives and to decide the pace of goal-directed actions) to enable correlation of behavior with neural activity. This allowed us to assess how self-generated actions arise from the contextual reorganization of mesocortical dynamics. We found that a distributed sequence of firing and oscillations in the VTA, PFC and OFC jointly set the goal of actions (i.e., deciding where to go), self-initiated the trial and determined the vigor and pace of the goal-directed action. This circuit sequence depended upon the reward context (deterministic or uncertain) and correlated with reward value and uncertainty, used to guide the animals' choices. Both cortical oscillations and distributed increase in firing emerged during learning as a reorganization of existing dynamics, with all the above-mentioned structures encoding prediction errors about the outcomes. Rather than fixed, this sequence could incorporate PFC oscillations or not, depending on the reward context, and the PFC could act in synergy or antagonistically to the VTA in co-determining action selection or execution.

A long-standing question in decision-making remains to identify subparts of the mesocorticolimbic circuits with computation underlying decisions. Many theories assume that goal-directed choice can be subdivided into a set of localized modules, each computing a variable of a phenomenological model of behavior (Cisek, 2012; Hunt and Hayden, 2017). For instance, in the influential reinforcement-learning framework, the phasic activity of dopamine neurons is classically viewed as calculating a reward prediction error (RPE, (Schultz et al., 1997)), i.e., the difference between actual and predicted value. In this framework, prefrontal and orbitofrontal cortices are implicated in value predictions, task state inference or beliefs (Lak et al., 2020; Sharpe et al., 2018; Starkweather et al., 2018). In stimulus-driven behaviors, frontal cortices are thus considered as upstream computations to dopamine RPE, inferring value or state predictions to be compared with actual outcomes by DA cells. In this framework, the frontal areas play an indirect role in behavior, by modulating a learning signal (Lak, Starkweather, Gershman & Uchida) or by informing areas that perform the choice, typically the basal ganglia (Klaus & Costa, Gershman & Uchida, Sharpe & Schoenbaum). By contrast, pharmacological and chemogenetic studies have shown an implication of prefrontal dopamine in decision-making, in particular under uncertainty (Onge and Floresco, 2010; Puig and Miller, 2012; Verharen et al., 2018), placing cortical dynamics downstream of dopamine modulation. This is in line with a vast literature assigning a more direct role to frontal cortices in ongoing action selection (Cai and Padoa-Schioppa, 2014; Cisek, 2012; Yoo and Hayden, 2018). The OFC and PFC have been shown to play crucial roles in reward-directed choice. In particular, neurons in the OFC have been proposed to encode the subjective value of options and to bias future choice (Jones et al., 2012; Kringelbach, 2005; Padoa-Schioppa and Conen, 2017; Takahashi et al., 2011)and neurons in the PFC have been suggested to compare the values of options in order to implement action policy (Cai and Padoa-Schioppa, 2014; Hunt and Hayden, 2017; Yoo and Hayden, 2018).

θ oscillations, widely observed across frontal cortical areas, are also implicated in general functions such as reward processing and goal-directed behavior (Amarante and Laubach, 2021; Cavanagh and Frank, 2014; Miller and Wallis, 2013; Puig and Miller, 2012).

Our recordings of neural activity in the VTA, the OFC and the PFC during the task revealed a distributed sequence of neural activity, which was consistent in each trial. This distributed mesocortical sequence started with an early phasic activity in the DA system together with an increased OFC firing, before the initiation of the action. This self-initiation was followed by a transition from increased spiking in the OFC to a θ oscillation. Increased OFC firing encoded the time of self-initiation, which would suggest an early cortical participation in decision-making. However, using optogenetic inhibition, we further show that VTA DA cells are likely to emit the "go" signal for initiation (see also below) as well as to favor the OFC transition from desynchronized spiking to θ oscillation. Our results thus suggest a recurrent OFC-VTA-OFC computation related to action initiation. We also found lines of evidence for a distributed computation of the overall pace of the action (from one rewarded location to another) among the PFC and VTA: firing in both structures correlated with the time to goal, which was also increased by the local inhibition of each structure. Our study thus lies with the alternative view that decisions may be distributed, i.e. formed in concert across the subparts of the mesocorticolimbic loop that perform parallel computations (Cisek, 2012; Hunt and Hayden, 2017; Steinmetz et al., 2019). The same view may hold for learning, as errors about the outcomes were present in all structures, during unexpected reward or omissions. Nevertheless, prediction errors were not similar: while VTA firing did conform with a reward prediction error (decreasing with omission and increasing with reward), the OFC signaled both (expected or unexpected) outcome regardless of its valence, and the PFC was engaged by unexpectedness only, in particular during omissions. Hence, for the learning part, our results are consistent with reinforcement-learning views that the VTA relates to value prediction errors (Dayan and Niv, 2008; Schultz et al., 1997) while the OFC is more dedicated in detecting state changes (Sharpe et al., 2018; Takahashi et al., 2009, 2011; Wilson et al., 2014) and the PFC with uncertainty (Starkweather et al., 2018) or deliberative strategies when the environment is uncertain (Bach and Dolan, 2012; Daw et al., 2005; Rushworth and Behrens, 2008; Rushworth et al., 2011). Such complementary predictions errors in VTA and frontal areas may arise from the same kind of comparisons (Hunt and Hayden, 2017) applied to different inputs, and might be needed for learning in complex settings (Sharpe et al., 2018; Stalnaker et al., 2015; Takahashi et al., 2009). More studies are thus needed to reconcile the modular and distributed views of learning and decision-making, with a potential important role for biophysical models to integrate neurophysiological findings at the circuit level (Hunt and Hayden, 2017).

However, predictions errors within the VTA, PFC and OFC were correlative, hence causality was not implied. Indeed, while the reinforcement-learning framework assigns a purely learning role to VTA RPEs (Coddington and Dudman, 2018; Sutton and Barto, 1998), experimental evidence has in fact been conflicting. On the one hand, mimicking RPE through optogenetic modulation of VTA DA phasic activity does affect learning in a manner predicted by the theory (Coddington and Dudman, 2018; Steinberg et al., 2013). On the other hand, early

phasic activity can be observed in DA neuron firing during uncued, self-initiated actions (Coddington and Dudman, 2018; Howe and Dombeck, 2016; Jin and Costa, 2010; Silva et al., 2018; Syed et al., 2015; Wassum et al., 2012). Such early phasic activity is compatible with the encoding of a reward prediction error occurring during the self-initiation of an action (Coddington and Dudman, 2018; Glimcher, 2011; Schultz, 2007; Sutton and Barto, 1998). For reinforcement models, DA cells would just be informed that action initiation is achieved elsewhere, this change in animal state being associated with an RPE that would not play any further role in ongoing behavior (Glimcher, 2011; Sutton and Barto, 1998). By contrast, recent studies have shown that phasic DA can also encode the kinetic variables of immediately occurring movements (Barter et al., 2015; Silva et al., 2018) and that modulation of phasic DA affects action initiation under some conditions (Fischbach-Weiss et al., 2018; Silva et al., 2018), but see (Coddington and Dudman, 2018). Hence the issue remains debated whether DA RPE at action initiation constitutes a learning signal for the next trials and/or a motivational command for the current trial.

In our recordings, we observed that early bursting in VTA DA neurons was not directly triggered by the MFB electrical stimulation. Indeed, it scaled with the reward probability, and it decreased at the time of the expected reward during reward omissions. All of these results comply with early bursting in VTA DA cells encoding RPE. However, early VTA DA firing also correlated with the vigor of the movement (estimated by the time to maximal speed and the time to goal). Furthermore, inhibiting this phasic activity reduced the time to goal by decreasing the maximal speed and delaying the action initiation. We thus causally implicate RPE-like VTA DA activity in the initiation and vigor of goal-directed actions (estimated by the time to maximal speed and the time to goal). Furthermore, inhibition of this phasic activity reduced the time to goal by decreasing the maximal speed and delaying the action initiation. We thus causally implicate RPE-resembling VTA DA activity in the initiation and vigor of goal-directed actions (Fischbach-Weiss et al., 2018; Silva et al., 2018), rather than a passive reflection of action initiation (Coddington and Dudman, 2018). As pDAn firing did not encode the time of initiation in advance, but temporary inhibiting VTA DA cells delayed initiation, we propose that VTA DA activity participates to a “go” command (Silva et al., 2018; Syed et al., 2015), with computations leading to action initiation performed elsewhere, the OFC in the present data, but also the striatum (Cisek, 2012; Klaus et al., 2016) . The encoding of dwell time by the OFC, as well as the transition from increased spiking to θ oscillations at action initiation, are in line with the state theory of the OFC (Sharpe et al., 2018; Wilson et al., 2014). In this theory, the OFC computes a “you are here” signal (within the task space), which is based on both external (cues) and inferred information. Action initiation would constitute a change in the animal’s state, associated with distinct OFC dynamics. This would also explain why OFC firing and oscillations react to any outcome (reward or omission), to signal changes between the different possible states of this task. However, in our task, action initiation is not caused by external cues signaling a state change, nor by the inference of a hidden state, such as for instance an unsignalled but imposed delay after which the action has to be initiated (Starkweather et al., 2018). On the opposite, we found that the OFC computed in advance the duration before state change (i.e. the dwell time), suggesting an active involvement of this structure in the change between states (a “you wait here” signal), rather than a passive monitoring role.

In line with a general role related to task space, we did not see any correlation between the discharge of OFC neurons, or the amplitude of θ oscillations, and the expected reward nor the reward uncertainty. This is at odds with accounts suggesting an OFC involvement in economic value (Cai and Padoa-Schioppa, 2014) or in computing confidence (i.e., the inverse of uncertainty, (Kepecs et al., 2008; O'Neill and Schultz, 2010) . However, value or uncertainty can be confounded with arousal and salience and causal involvement of OFC in economic choice is lacking (Stalnaker et al., 2015) . In contrast with the OFC, we found clear correlates of value in the VTA DA firing, and of uncertainty in the PFC firing. Quantitative encoding of probabilistic reward complies with RPE theories and has been described at length (Dabney et al., 2020; Eshel et al., 2016; Tobler et al., 2005). PFC δ oscillations have also been implicated in motivation (Knyazev, 2012). However, expected uncertainty was not aggregated with expected value in PFC activity, contrary to observations in humans (Lebreton et al., 2009; Tobler et al., 2009). This might relate to task differences rather than species differences. Indeed, the PFC has been implicated in selecting the strategy for goal-directed choice (Miller and Wallis, 2013; Sharpe et al., 2018). In our task, we did not observe any increased firing nor any δ oscillations in the PFC in the deterministic context, suggesting that the PFC is mostly needed in the probabilistic context, i.e. for decisions under uncertainty (Bach and Dolan, 2012; Rushworth and Behrens, 2008). As mice used uncertainty to guide their decisions in the probabilistic context, the PFC may encode uncertainty in our task in order to represent decision-guiding rules, rather than uncertainty-modulated choice values per se.

Inhibiting the PFC by a local infusion of muscimol led to an increase of uncertainty-seeking, showing that the encoding of uncertainty by the PFC had an inhibitory influence on uncertainty-biased choices. As dopamine generally exerts a positive influence on uncertainty-seeking (Naudé et al., 2016; Onge and Floresco, 2010), this suggests antagonistic influences of VTA and PFC on the motivation induced by reward uncertainty. This might reflect a difference between how model (or belief)-based and model-free control may treat uncertainty. In the probabilistic context, the uncertainty associated with the reward is known by the animals, i.e. it is a form of expected uncertainty (Yu and Dayan, 2005). In the simplest form of curiosity, expected uncertainty or variability may exert a positive motivational influence in the form of a bonus added to the expected value by DA cells, in order to promote the steady-state exploration of unpredictable options (Anselme et al., 2013; Naudé et al., 2016). By contrast, in deliberative strategies putatively implicating the PFC, the known uncertainty may be treated as uninformative noise that has to be discarded from the decision strategy (Yu and Dayan, 2005).

Strikingly, the VTA and PFC had opposite influences on uncertainty-related choices, but synergistic influences on the pace of actions. Suppression of the early phasic activity of the pDA neurons and PFC inhibition induced similar effects on the pace of behavior, with an increase in time to goal and a decrease in total transition number. This could either suggest a sequential computation (e.g. DA-then-cortex), a recurrent one (DA-cortex-DA, as suggested above for the VTA and OFC), or two independent processes causally linked to movement initiation. More work is thus needed to dissect PFC-DA interactions (Ellwood, Jo, Park) in ongoing, self-generated decisions. We however suggest that this circuit presents a "fluid" organization as described for invertebrate circuits (Marder

and Bucher, 2007) : depending on the reward context, the PFC may flexibly integrate the basic circuit formed by the VTA-OFC loop in charge for the initiation of a goal-directed action, adding uncertainty-based computations to the distributed sequence. On the same order of ideas, the emergence of these activities with learning followed a reorganization of existing circuit dynamics. Indeed, completion of learning induced the emission of bursts by dopamine cells at each trial, but also corresponded to an increase in the number of trials. The combination of these two features would have suggested an overall increase in pDAn firing frequency, but we did not observe any change. Hence early bursting in dopamine neurons did not rely on additional spikes, but rather on a dynamical re-organization towards synchronous bursting activity. This increased in early DA activity correlated also with a locking of θ oscillations in the OFC rather than an increase in θ power. This suggests that the characteristics of the VTA-OFC-PFC we and others unraveled, i.e. distributed yet distinct contributions to both learning and decisions, may rely on the alignment of subparts' dynamics at relevant timings, rather than on increases in activities of separate modules each computing a decision variable. While this alignment of mesocortical dynamics is usually forced by a stimulus, we show here that it reorganizes even without stimulus, resulting in the self-generation of goal-directed actions.

Acknowledgements:

We are grateful to the animal facilities (IBPS), Camille Robert and Paris Vision Institute AAV production facility for viral production and purification. This work was supported by the Centre National de la Recherche Scientifique CNRS UMR 8246, INSERM U1130, the Foundation for Medical Research (FRM, Equipe FRM DEQ2013326488 to P.F), the French National Cancer Institute Grant TABAC-16-022 et TABAC-19-020 (to P.F.), French state funds managed by the ANR (ANR-16 Nicostress to PF).

Author contributions:

JN and PF designed the study. EB SD ST CPS MC JN performed electrophysiological recordings. EB YAT CPS performed the behavioral experiments. SD and ST contributed to setup developments. EB JN SD ST CPS performed the surgeries and virus injections. EB performed the immunohistochemistry. DD provided the viruses. LT provided the DAT-Cre mice. AM developed the optogenetics setup. JN and PF developed the model. EB, JN and PF analyzed the data. EB, JN and PF wrote the manuscript with inputs from the other authors.

Declaration of interests:

The authors declare no competing financial interests.

Methods

Animals

Experiments were performed on adult C57Bl/6Rj DAT^{iCRE} and Wild-Type (Janvier Labs, France) mice. Male mice, from 8 to 16 weeks old, weighing 25-35 grams, were used for all the experiments. They were kept in an animal facility where temperature (20 ± 2°C) and humidity were automatically monitored and a circadian light cycle of 12/12-h light-dark cycle was maintained. All experiments were performed in accordance with the recommendations for animal experiments issued by the European Commission directives 219/1990, 220/1990 and 2010/63, and approved by Sorbonne University.

AAV production

AAV vectors were produced as previously described using the cotransfection method and purified by iodixanol gradient ultracentrifugation[51]. AAV vector stocks were tittered by quantitative PCR (qPCR)[52] using SYBR Green (Thermo Fischer Scientific).

Intracranial self-stimulation electrode and recording electrode implantation

Mice were anaesthetized with a gas mixture of oxygen (1 L/min) and 1-3 % of isoflurane (Piramal Healthcare, UK), then placed into a stereotaxic frame (Kopf Instruments, CA, USA). After the administration of a local anesthetic (Lurocain, 0.1 mL at 0.67 mg/kg), a median incision revealed the skull which was drilled at the level of the Median Forebrain Bundle (MFB), the OFC, the PFC or the VTA. Dental cement (SuperBond, Sun Medical) was used to fix the implant to the skull. A bipolar stimulating electrode for ICSS was then implanted unilaterally (randomized) in the brain (stereotaxic coordinates from bregma according to mouse after Paxinos atlas: AP -1.4 mm, ML ±1.2 mm, DV -4.8 mm from the brain). Bipolar recording electrodes were implanted in the lateral OFC (AP +2.6 mm, ML ±1.5 mm, DV -1.7 mm from the brain) and the medial PFC (AP +1.65 mm, ML ±0.5 mm, DV -1.8 mm from the brain). Octrodes recording electrodes were implanted in the VTA (AP -3.15 to -3.25 mm, ML ±0.5 mm, DV -4.1 to 4.25 mm from the brain). After stitching and administration of a dermal antiseptic, mice were then placed back in their home-cage and had, at least, 5 days to recover from surgery. An analgesic, buprenorphine solution at 0,015 mg/L (0,1 mL / 10 g), was delivered after the surgery and if necessary, the following recovering days. The efficacy of electrical stimulation was verified through the rate of acquisition during the deterministic setting (see behavioral methods).

Virus injections

DAT^{iCRE} mice were anaesthetized (Isoflurane 1-3%) and were injected unilaterally (randomized left/right side and ipsi/contralateral side) in the VTA (1 µL, coordinates from bregma: AP -3.15 to -3.25 mm; ML ±0.5 mm; DV -4.55 mm from the skull) with an adeno-associated virus (AAV5.EF1α.DIO.Jaws.eGFP 1.16e^{13} ng/µL or AAV5.EF1α.DIO.YFP 6.89e^{13} or 9.10e^{13} ng/µL). A double-floxed inverse open reading frame (DIO) allowed to restrain the expression of Jaws (red-shifted cruxhalorhodopsin) to VTA dopaminergic neurons.

Polyelectrodes:

Hand-made poly-electrodes (bundle of 8 electrodes, “octrodes”) were obtained by twisting eight polyimide-insulated 17 µm Nickel-Chrome wires. The use of eight channels relatively close together allows for a better discrimination of the different neurons. Before implantation and recording, the octrodes were cut at suitable length and plated using a Platinium-PEG solution to lower their impedance to 150-400 KOhms and improve the signal-to-noise ratio. The free ends of the octrodes were connected to the holes of EIB-18 (electrode interface board, Neuralynx) and fixed with pins. We manufactured a microdrive system (home-made 3D conception and printing) consisting of a main body, on which is mounted the EIB, and a driving screw, with a sliding part design to contain the two octrodes. This microdrive allowed moving through the VTA in order to sample neuronal populations.

Bipolar electrodes:

Hand-made bipolar electrodes were obtained by twisting two Teflon-insulated (60 µm) Stainless Steel wires. Two configurations were used. For the first one, the tips of the bipolar electrodes were cut so that they are spaced of less than 0.5 mm apart. For the second one, the reference tip was wound around the recording one, at a distance of less than 0.5 mm from the recording endpoint. These electrodes are designed so the two tips are oriented perpendicular to the dipoles formed by cortical pyramidal neurons. The first configuration was used for OFC recording electrodes, and the second one for PFC recording electrodes. IntraCranial Self-Stimulation (ICSS) electrodes were made as the second configuration with a 80 µm Stainless Steel wire. Bipolar electrodes were connected to the EIB during the surgery, by fixing the free ends with pins.

Immunochemistry

After euthanasia, brains were rapidly removed and fixed in 4% paraformaldehyde (PFA). After a period of at least three days of fixation at 4°C, serial 60-µm sections were cut with a vibratome (Leica). Immunostaining experiments were performed as follows: VTA brain sections were incubated for 1 hour at 4°C in a blocking solution of phosphate-buffered saline (PBS) containing 3% bovine serum albumin (BSA, Sigma; A4503) (vol/vol) and 0.2% Triton X-100 (vol/vol), and then incubated overnight at 4 °C with a mouse anti-tyrosine hydroxylase antibody (anti-TH, Sigma, T1299) at 1:500 dilution, in PBS containing 1.5% BSA and 0.2% Triton X-100. The following day, sections were rinsed with PBS, and then incubated for 3 hours at 22-25 °C with Cy3-conjugated anti-mouse and secondary antibodies (Jackson ImmunoResearch, 715-165-150) at 1:500 in a solution of 1.5% BSA in PBS, respectively. After three rinses in PBS, slices were wet-mounted using Prolong Gold Antifade Reagent (Invitrogen, P36930). Microscopy was carried out with a fluorescent microscope, and images captured using a camera and analyzed with ImageJ. In the case of optogenetic experiments on DAT^{iCRE} mice, identification of the transfected neurons by immunohistofluorescence was performed as described above, with the addition of 1:500 Chicken-anti-GFP primary IgG (ab13970, Abcam) in the solution. A Goat-anti-chicken AlexaFluor 488 (1:500, Life Technologies) was then used as secondary IgG. Neurons labelled for TH in the VTA allowed to confirm their neurochemical phenotype, and those labelled for GFP to confirm the transfection success

Intracranial self-stimulation (ICSS) bandit task

Behavioral set up: The ICSS bandit task took place in a circular open field with a diameter of 67 cm. Three explicit square-shaped marks (1x1 cm) were placed in the open field, forming an equilateral triangle (side = 35 cm). Entry in the circular zones (diameter = 6 cm) around each mark was associated with the delivery of a rewarding ICSS stimulation. Experiments were performed using a video camera, connected to a video-tracking system, out of sight of the experimenter. A LabVIEW (National Instruments) application precisely tracked and recorded the animal's position with a camera (20 frames/s). When a mouse was detected in one of the circular rewarding zones, an electrical stimulator received a TTL signal from the software application and generated a 200 ms-train of 0.5-ms biphasic square waves pulsed at 100 Hz (20 pulses per train). ICSS intensity was adjusted, within a range of 20 to 200 µA, during training (see training settings) and then kept constant, so that mice would achieve between 50 and 150 visits per session (5min duration) for two successive sessions, and then kept constant for all the experiment. Mice with insufficient scores in the PS and DS (<40 visits despite increasing the maximum intensity of 200 µA) were excluded.

Baseline behavior: Prior to the ICSS bandit task, three control sessions were performed. First, spontaneous neuronal activity was recorded in the mice home-cages for 10 minutes. Second, neuronal activity was recorded while random ICSS were delivered to the mice in its home-cage, to assess the direct effect of the stimulation onto neuronal activity. Third, behavioral and neuronal activity were recorded for 30 min, while the mice were exploring the open-field for the first time ("habituation", without the presence of the three rewarding locations).

Training setting: The training consisted of two settings: the deterministic setting (DS) and the probabilistic setting (PS), consisting of 10 daily sessions of 5 min for the DS and 10 min for the PS. In the DS, all zones were associated with an ICSS delivery (P = 100%). However, two consecutive rewards could not be delivered on the same target, which motivates mice to alternate between targets. In the PS, the zones were associated with three different probabilities (P = 25%, P = 50%, P = 100%) to obtain an ICSS stimulation. The probabilities' locations were pseudo-randomly assigned per mouse.

Data acquisition per experimental group: For optogenetics experiments, the DAT^{iCRE} mice (n = 16) completed the training, followed by a schedule of 4 days of paired sessions with photo-stimulation (ON) alternated with days without photostimulation (OFF). The averages of the ON and OFF days were compared in a paired manner.

Optogenetics experiments: For optogenetic experiments on freely moving mice, an optical fiber (200 µm core, NA = 0.39, Thor Labs) coupled to a ferule (1.25 mm) was implanted just above the VTA ipsilateral to the viral injection (coordinates from bregma: AP -3.1 mm, ML ±0.5 mm, DV 4.4 mm), and fixed to the skull with dental cement (SuperBond, Sun Medical). The behavioral task began at least 4 weeks after virus injection to allow the transgene to be expressed in the target dopamine cells. An ultra-high-power LED (520 nm, Prizmatix) coupled to a patch cord

(500 µm core, NA = 0.5, Prizmatix) was used for optical stimulation (output intensity of 10 mW). Optical stimulation during the behavioral experiment was continuously delivered for 500 ms, starting 500 ms after animal's detection in a location. The ON and OFF schedule (OFF-ON-OFF-ON-OFF) was following the last week of deterministic training. The optical stimulation cable was plugged onto the ferrule during 5 experimental sessions to prepare the animals and control for latent experimental effects.

Intracranial injections of muscimol: A solution of of muscimol (TOCRIS) (0.5µg/µL) was infused in the PFC over 20-30 minutes before the beginning of the ICSS bandit task experiment. The bilateral infusion of 0.4µL was performed at a rate of 0.2µL/min using a double injector (Univentor). Before each experiment session, a double injection cannula (2.5 mm, 0.5 mm projection) was inserted into the implanted bilateral cannula guide (length below pedestal 2.5 mm). The injection cannula was connected to a multi-syringe pump (Univentor) that allowed saline or muscimol injection. The saline and muscimol schedule (saline-muscimol-rest-saline-muscimol) was following the last week of probabilistic training. The injection system was plugged onto the canula guide before 5 experimental sessions to prepare the animals and control for latent experimental effects.

Behavioral measures: For all groups of mice, the trajectory was smoothed using a triangular filter allowing the determination of speed profile, which corresponds to instantaneous speed as a function of time, and time of maximal speed within a trial. The following measures were analyzed in the DS and compared in the PS, as well as in the DS for the OFF vs ON Jaws experiment, or in the PS for the Sal vs Mus experiment: i) number of visits, ii) time-to-goal, iii) choice repartition (proportion of visits p_{25}, p_{50} and p_{100}), iv) percentage of directional changes (n^{th} visit = n^{th} visit+2). Furthermore, the ICSS bandit task can be seen as a Markovian decision process. Every transition between zones can be considered as a binary choice between two probabilities, since the occupied zone cannot be reinforced twice in a row. The sequence of choices per session is summarized by the proportional result of the sum of three specific binary choices (or gambles, i.e., total visits zone 1/total visits zone 1+2). The three gambles (G) were named after the point on which the mouse is positioned at the time of the choice: G_{25} = 100 % vs 50 %, G_{100} = 50 % vs 25 % and G_{50} = 100 % vs 25 %. The target choice in these gambles reflects the balance between exploitative (choosing the most valuable option) and exploratory (choosing the least valuable option) choices. With a softmax based decision-making model fitted in the laboratory, we computed three parameters: the value sensitivity or inverse temperature (the power to discriminate between values in a binary choice), the uncertainty bonus (the preference for expected uncertainty, considering the reward variance of every option in a binary choice) and the motor cost to do a directional change (a decrease in the target value if it requires to go back to the previous target).

Modeling: Decision-making models determined the probability P_i of choosing the next state i, as a function (the "choice rule") of a "decision variable". Because mice could not return to the same rewarding target, they had to choose between the two remaining ones. Accordingly, we modeled decisions between two alternatives labelled A and B and used a softmax choice rule defined by $P_A = 1 / (1+e^{-ß(vA-vB)})$ where β is an inverse temperature parameter

reflecting the sensitivity of choice to the difference between decision variables and V_i the value of an option. The value V of an option is modelled as the expected (average) reward + expected uncertainty + U-turn cost [16,30]. This compound value is then nested in the softmax choice rule, given a 6*3 matrix that described the probability of a choice between A, B and C (the three targets) depending on the two previous choices. As an example, in the probability to choose (A, B, C) after performing the sequence BA, the value is given by $(0, p_b + \varphi p_b*(1-p_b) - \kappa, p_c + \varphi * p_c*(1-p_c))$ while after the sequence CA the value is given by $(0, p_b + \varphi * p_b*(1-p_b), p_c + \varphi * p_c*(1-p_c) - \kappa)$ (same for AB, CB and AC, BC). The free parameters of the model were fitted by maximizing the data likelihood. Given a sequence of choice $c = c_{1..T}$, data likelihood is the product of their probability (given by Equation 1). We used the *optim* function in R or *fmincon* function in Matlab to perform the fits, with the constraints that $\beta \in]0,10]$, $\varphi \in]-1,5]$ and $\kappa \in]0,5]$.

Statistical analysis: All statistical analyses were computed using Matlab and Python with custom programs. Results were plotted as a mean ± s.e.m. The total number (n) of observations in each group and the statistics used are indicated in figure legends. Classical comparisons between means were performed using parametric tests (Student's T-test, or ANOVA for comparing more than two groups) when parameters followed a normal distribution (Shapiro test $P > 0.05$), and non-parametric tests (here, Wilcoxon or Mann-Whitney) when the distribution was skewed. Multiple comparisons were Bonferroni corrected. Probability distributions were compared using the Kolmogorov–Smirnov (KS) test, and proportions were evaluated using a chi-squared test (χ^2).

Electrophysiological recordings

All extracellular potentials recordings were performed using a digital acquisition system (Digital Lynx SX; Neuralynx) together with the Cheetah software. Broadband signals from each wire were filtered between 0.1 and 9000 Hz and recorded continuously at 32 kHz.

Multi-unit activity recordings: To extract spike timing, signals were band-pass-filtered between 600 and 6000 Hz and sorted offline. Spike clustering was cross-validated by using both SpikeSort3D (Neuralynx) and custom-written Matlab (The Mathworks) routines. The electrophysiological characteristics of VTA neurons were analyzed in the active cells encountered by systematically moving down the bi-octrode.

Local-field potential recordings: To extract low-frequency variations of extracellular potential, signals were low-pass-filtered below 300 Hz.

Population firing: To extract spike timing of the neuronal population, signals were band-pass-filtered between 600 and 6000 Hz and sorted offline. Because population firing originates from bipolar electrodes with only one recording wire, no clustering could be considered.

Electrophysiological data analysis

Identification of DA cells: Extracellular identification of putative DA neurons (pDAn) was based on their location as well as on a set of unique electrophysiological properties that characterize these cells in vivo: 1) a typical triphasic action potential with a marked negative deflection; 2) a characteristic long duration (>2.0 ms) action potential; 3) an

action potential width from start to negative trough >1.1 ms; 4) a slow firing rate (<12 Hz) with an irregular single spiking pattern and occasional short, slow bursting activity. Putative GABA neurons were characterized by a characteristic short duration of action potential from start to negative trough (<1.0 ms), and a high firing rate (>12Hz). D2 receptors (D2R) pharmacology was also used for confirming the DA neurons identification: after a baseline periode (5 min) and a saline (10 min) injection, quinpirole (1mg/kg, D2R antagonist) was injected (30 min recording), followed by an eticlopride (D2R agonist) injection (1mg/kg, 10 min recording). Since most DA, but not GABA neurons, express inhibitory D2 auto-receptors, neurons were considered as pDA neurons if quinpirole induced at least 30% decrease in their firing rate, while eticlopride restored firing above the baseline. Nevertheless, as continuous D2 pharmacology could have affected both baseline DA neurons firing and decision-making (Fobbs and Mizumori, 2014), we allowed the mice to recover two days after this experiment. We thus performed pharmacological confirmation (1) when first encountering a putative DA neuron in a given mouse or (2) at the end of the week if at least one putative neuron was present during the behavioral experiment. Neurons were considered DA only if they responded to the pharmacology, or if they presented electrophysiological characteristics defined above and were recorded between two positive pharmacological experiments.

Firing analysis: Spontaneous DA cell firing was analyzed with respect to the average firing rate and the percentage of spikes within bursts (%SWB, number of spikes within bursts, divided by total number of spikes). Bursts were identified as discrete events consisting of a sequence of spikes such that: their onset is defined by two consecutive spikes within an interval <80 ms and they terminated with an interval >160 ms. Phasic activity is defined as spikes falling into bursts, while tonic activity comprises spikes outside bursts. Peri-event time histograms were constructed based on 1 ms-bins rasters, convolved with a Gaussian kernel corresponding to the same event.

Wavelet analysis: Because extracellular field potentials (EFP) are non-stationary signals, they are transformed offline using a Morlet wavelet transform. This process is defined as the convolution product between the EFP signal and dilated forms of wavelets normalized to 1 (Quyen and Bragin, 2007).

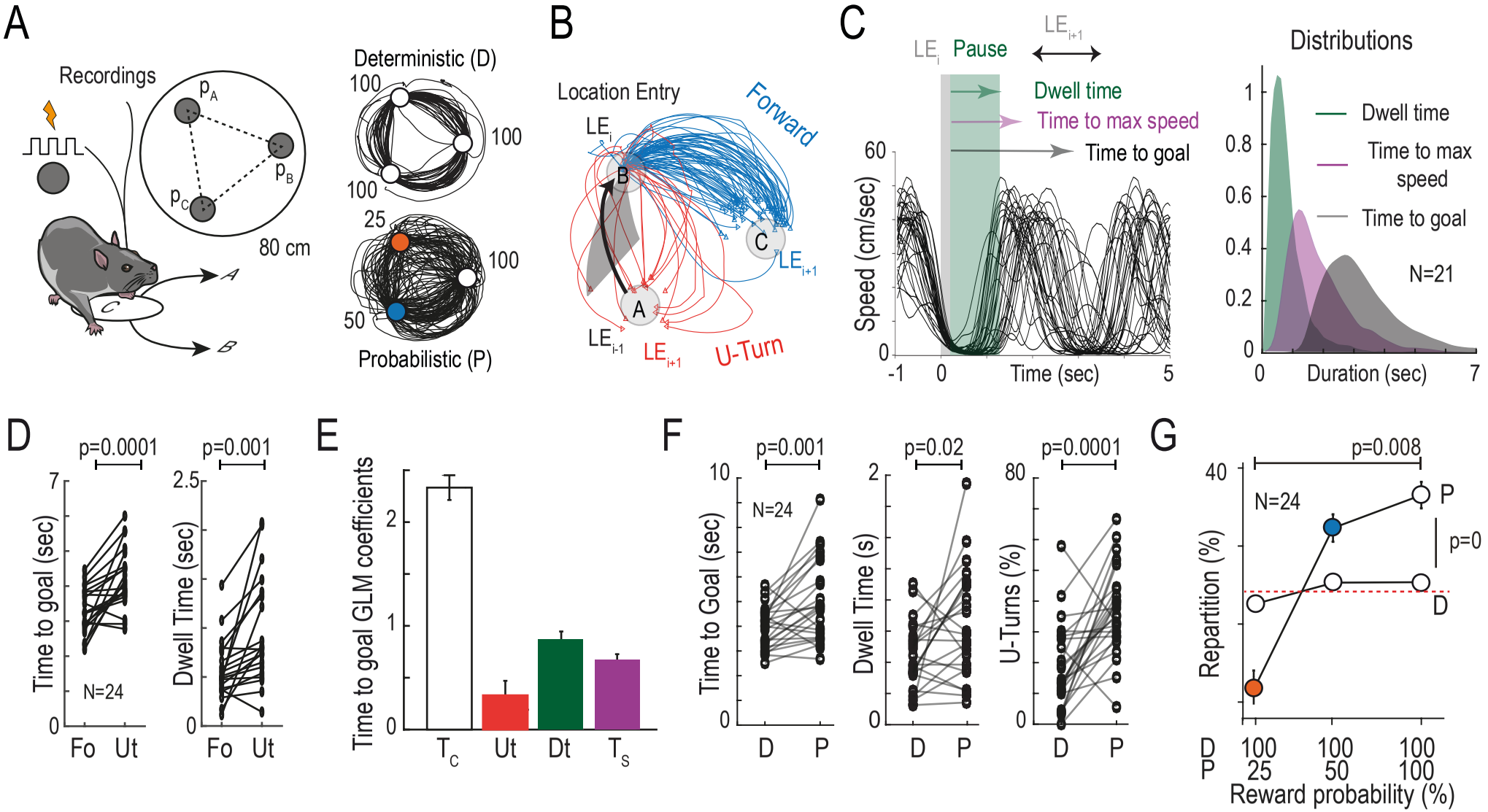

A
Recordings
p_A
p_B
p_C
80 cm
A
B
C
Deterministic (D)
100
100
100
25
100
50
Probabilistic (P)
B
Location Entry
Forward
LE_i
B
C
LE_{i+1}
A
LE_{i-1}
LE_{i+1}
U-Turn
C
LE_i
Pause
LE_{i+1}
Dwell time
Time to max speed
Time to goal
Speed (cm/sec)
60
0
-1
0
Time (sec)
5
Distributions
1
0.8
0.6
0.4
0.2
0
Dwell time
Time to max speed
Time to goal
N=21
0
Duration (sec)
7
D
p=0.0001
p=0.001
Time to goal (sec)
7
0
N=24
Fo
Ut
Dwell Time (sec)
2.5
0
Fo
Ut
E
Time to goal GLM coefficients
2
1
0
T_C
Ut
Dt
T_S
F
p=0.001
p=0.02
p=0.0001
Time to Goal (sec)
10
0
N=24
D
P
Dwell Time (s)
2
0
D
P
U-Turns (%)
80
0
D
P
G
p=0.008
Repartition (%)
40
N=24
P
p=0
D
D 100 100 100
P 25 50 100
Reward probability (%)

Figure 1

Figure 1: Self-paced decisions in a mouse spatial task based on intracranial self-stimulations. (A) Left: illustration of the task design. Three explicit square locations were placed in the open field (0.68-m diameter), forming an equilateral triangle (35-cm side). Each location is associated with a given probability of intracranial self-stimulation (ICSS) delivery (P = 100%, 50%, 25%) when the mouse is detected in the location area. Animals could not receive two consecutive ICSS at the same location. Right: example trajectories (5 min) showing that mice alternated between rewarding locations in the deterministic (D) and probabilistic (P). **(B)** Animals vary between forward trajectories, in which mice keep the direction of their last choice, thus performing 'A-B-C' sequences; and U-turn trajectories, in which mice go backward after to their previous location, corresponding to A-B-A sequences. **(C)** Left: examples of instantaneous speed showing that the animals almost stopped at the time of the reward after location entry (henceforth "location entry" LE), stayed immobile for a short dwelling period, then accelerated toward their next location. These bouts of activity can be described using three observables, the dwell time (time from ICSS delivery during which speed is less than 10 cm/s), the time to the maximal speed, and the total time to goal from one location to the next. Right: distribution for these three parameters, for all trials of all mice. **(D)** Time to goal and dwell time were higher for mice performing U-turn trajectories compared to Forward trajectories in the D setting (paired student t.test,n=24;$p<0.001$). **(E)** Coefficients of the generalized linear model of time to goal in the D setting: constant term (Tc), U-turn (categorical variable, either U-turn or forward), dwell time (Dt) and time to maximal speed (Ts). **(F)** Time to goal, dwell time, and proportion of U-turns increased in the probabilistic (P) setting compared to the deterministic (D) setting (paired student t.test,n=24;$p<0.001$). **(G)** Proportion of choices of the three rewarding locations, as a function of reward probability in the P and D setting (N=24).

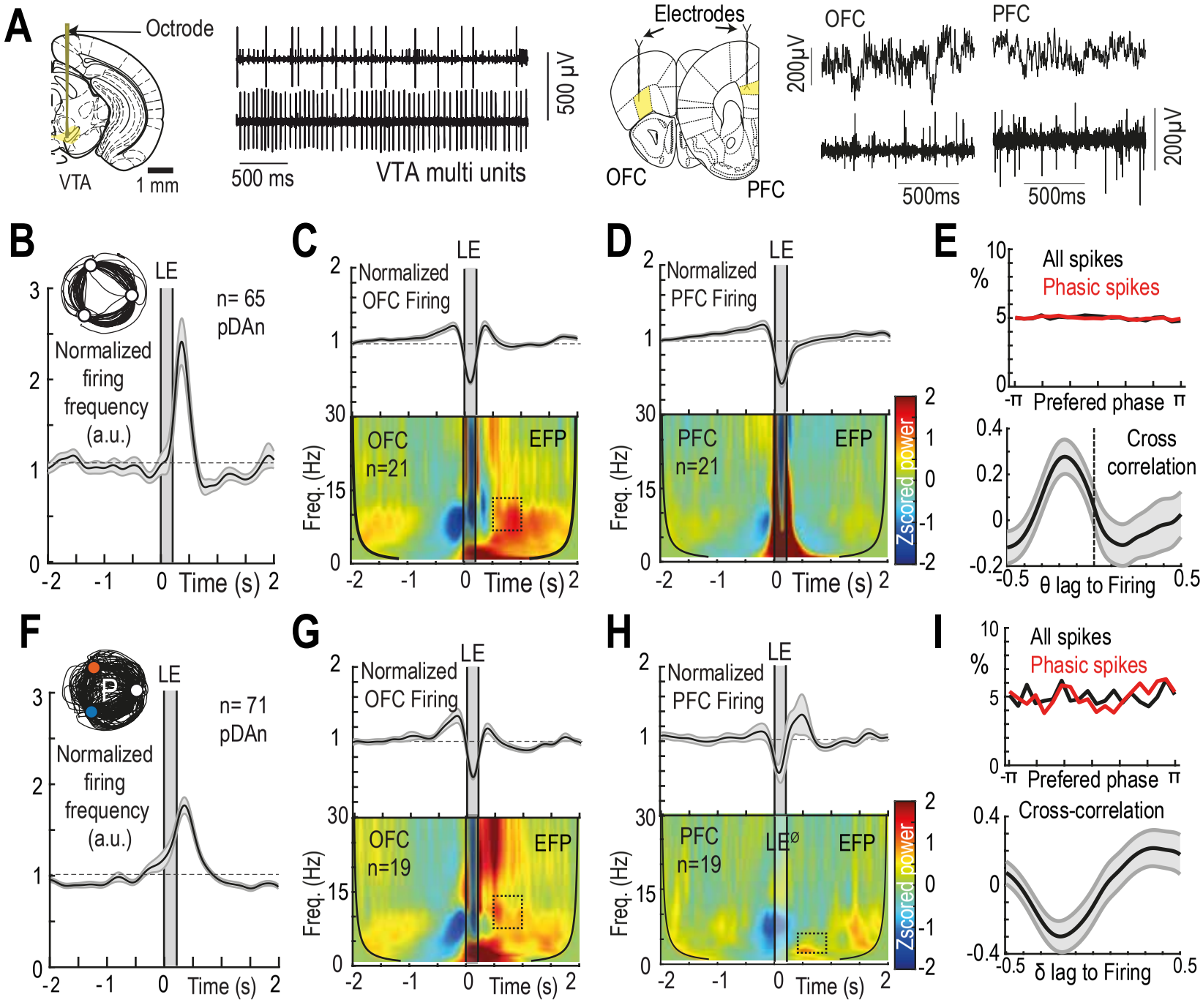

A
Octrode
VTA
1 mm
500 ms
VTA multi units
500 µV
Electrodes
OFC
PFC
200µV
500ms
B
LE
n= 65
pDAn
Normalized
firing
frequency
(a.u.)
Time (s)
C
Normalized
OFC Firing
OFC
n=21
EFP
Freq. (Hz)
D
Normalized
PFC Firing
PFC
n=21
Zscored power
E
All spikes
Phasic spikes
Prefered phase
Cross
correlation
θ lag to Firing
F
n= 71
pDAn
G
OFC
n=19
H
PFC
n=19
I
Cross-correlation
δ lag to Firing

Figure 2

Figure 2: VTA, PFC and OFC during deterministic and probabilistic reward contexts.

(A) Left: filtered (600-6000 Hz) extracellular recordings in the VTA showing multiple unit activity. Right: Extracellular recordings in the OFC and PFC showing extracellular field potential (filtered 0.1-300 Hz) (top) or population activity (filtered 600-6000 Hz) (bottom). **(B)** Normalized firing frequency from VTA pDAn around the time at which the animal is first detected in the location (LE), i.e. reward delivery in the D setting. **(C)** Normalized firing frequency from OFC population (top) and time-resolved power spectral density (PSD; using a complex Morlet wavelet transform) of OFC EFP (bottom) around the time at which the animal is first detected in the location, (i.e. reward delivery in the D setting). PSD is Zscored over the 2 sec period preceding the location entry (LE). **(D)** Normalized firing frequency from PFC population (top) and time-resolved power spectral density (PSD; using a complex Morlet wavelet transform) of PFC EFP (bottom) around the time at which the animal is first detected in the location, (i.e. reward delivery in the D setting). PSD is Zscored over the 22 sec period preceding the location entry (LE). **(E)** Top: Proportion of OFC population activity locked onto the phase of OFC θ-band frequency (7-14 Hz), for all spikes (black) or only spikes occurring early in the trial ('phasic', red). Bottom: Trial by trial cross-correlation between OFC population firing activity and the θ power in OFC. **(F,G,H)** same as (B,C,D) but in the P setting. The PFC PSD showed in (H) is centered around reward omissions to prevent the ICSS artefact from obscuring the low frequency power. **(I)** Top: Proportion of PFC population activity locked onto the phase of PFC δ-band frequency (3-6 Hz), for all spikes (black) or only spikes occurring early in the trial ('phasic', red). Bottom: Trial by trial cross-correlation between PFC population firing activity and the δ power in OFC.

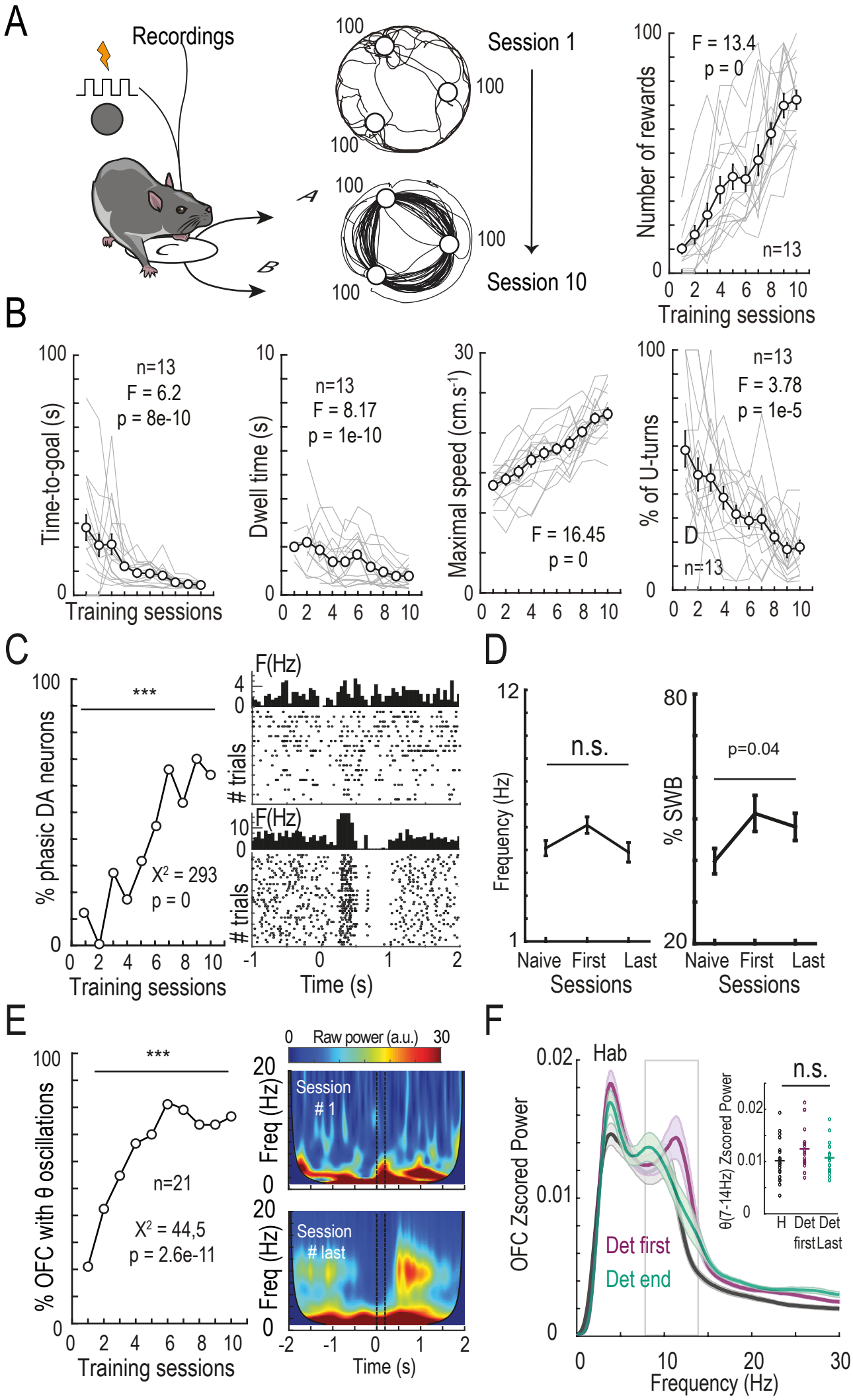

A
Recordings
Session 1
Session 10
Number of rewards
F = 13.4
p = 0
n=13
Training sessions
B
Time-to-goal (s)
n=13
F = 6.2
p = 8e-10
Dwell time (s)
n=13
F = 8.17
p = 1e-10
Maximal speed (cm.s-1)
F = 16.45
p = 0
% of U-turns
n=13
F = 3.78
p = 1e-5
C
% phasic DA neurons
***
X² = 293
p = 0
F(Hz)
# trials
Time (s)
D
Frequency (Hz)
n.s.
% SWB
p=0.04
Naive First Last
Sessions
E
% OFC with θ oscillations
***
n=21
X² = 44,5
p = 2.6e-11
Raw power (a.u.)
Session # 1
Session # last
Freq (Hz)
F
Hab
OFC Zscored Power
Det first
Det end
θ(7-14Hz) Zscored Power
H Det first Det Last

Figure 3

Figure 3: Early VTA and OFC activities emerge with learning. (A) Modifications of the number of rewards along the learning in the D setting. Number of rewards were evaluated from session 1 (example trajectory middle above) to session 10 (example trajectory middle below). (Right) Dots and vertical bar are mean ±sem. Grey lines indicate modifications of the number of rewards per individual (n=13 animals) (Anova, $p<0.001$) **(B)** Same as A for from right to left: time to goal, dwell time, maximal speed, and proportion of U-turns along the learning sessions. **(C)** (Left) Proportion of pDAn modulated in-between two locations, along learning sessions (N=10). (Right) Examples of raster plots, centered on location entry, for a VTA pDAn early in learning (top) and another at the end of the learning (bottom) **(D)** Average firing frequency (left) and % spikes within bursts (right) throughout learning (naive: open-field without ICSS, first: sessions 1-5, last: sessions 6-10). **(E)** Proportion of OFC θ power modulated in-between two locations, along learning sessions. (Right) Examples of OFC time-resolved power spectral density centered on location entry, for one OFC early in learning (top) and for the same example OFC at the end of the learning (bottom). **(F)** Mean Zscored power of Fourrier transform spectra of OFC during open field habituation (Hab, black, N=21, see Methods), early in learning (Det first, purple, N=20), at the end of the learning (Det end, green, N=19). Power in the θ-band frequency (7-14 Hz, grey box) shows no significant differences between the three conditions (t-test, ns).

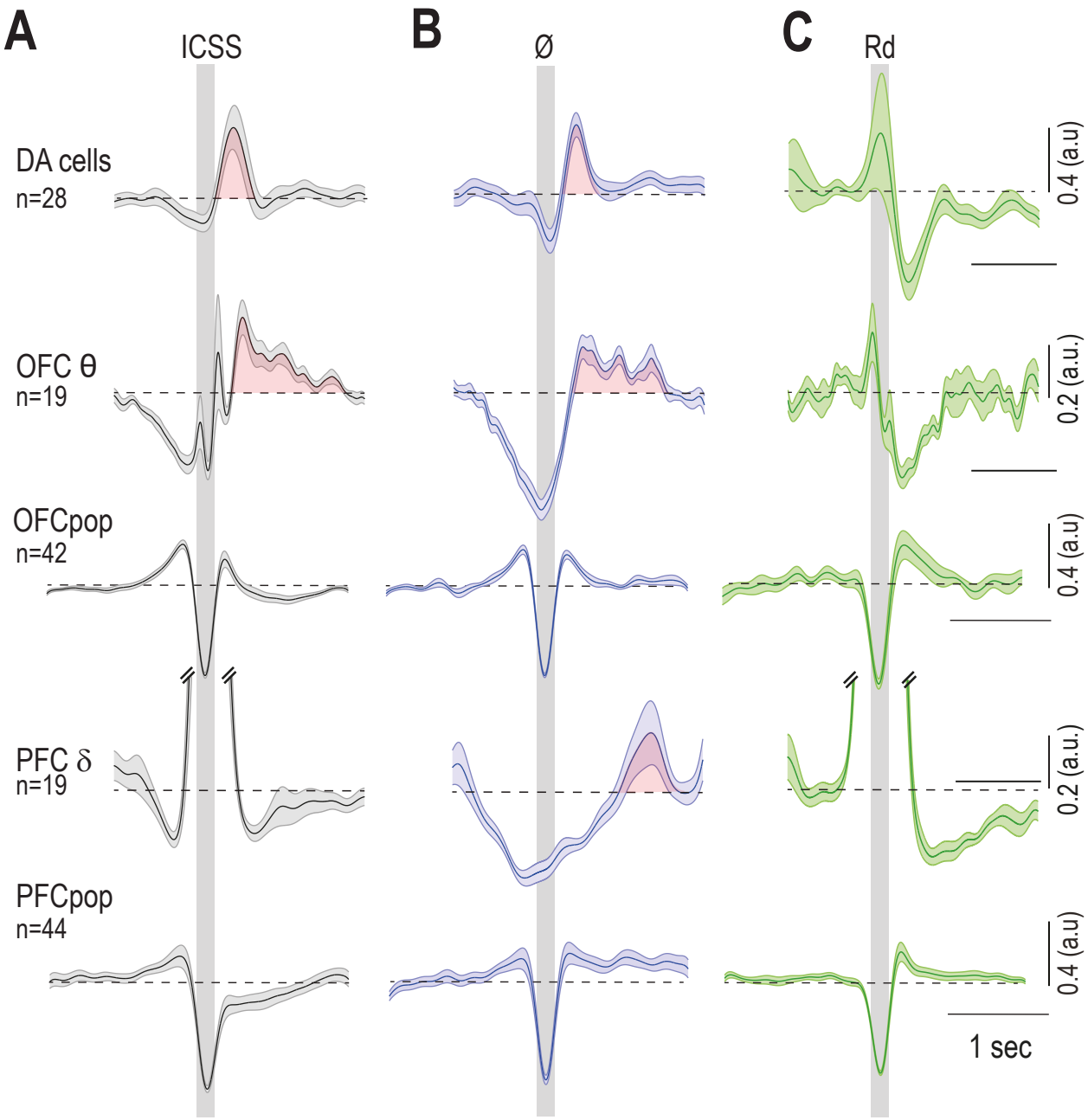

A
ICSS
B
Ø
C
Rd
DA cells
n=28
OFC θ
n=19
OFCpop
n=42
PFC δ
n=19
PFCpop
n=44
0.4 (a.u)
0.2 (a.u.)
0.4 (a.u)
0.2 (a.u.)
0.4 (a.u)
1 sec

Figure 4

Figure 4: VTA, OFC and PFC activities following expected reward, unexpected omission and unexpected reward. (A) From top to bottom: VTA pDAn normalized firing, OFC θ oscillation power, OFC normalized population firing, PFC δ oscillation power, and PFC normalized population firing, centered on expected reward delivery upon location entry at the beginning of the P setting. **(B)** Same as (A), centered on unexpected omission of reward delivery upon location entry at the beginning of the P setting. **(C)** Same as (A), centered on unexpected reward delivery upon random intra-cranial stimulation in the home-cage, before the beginning of the conditioning.

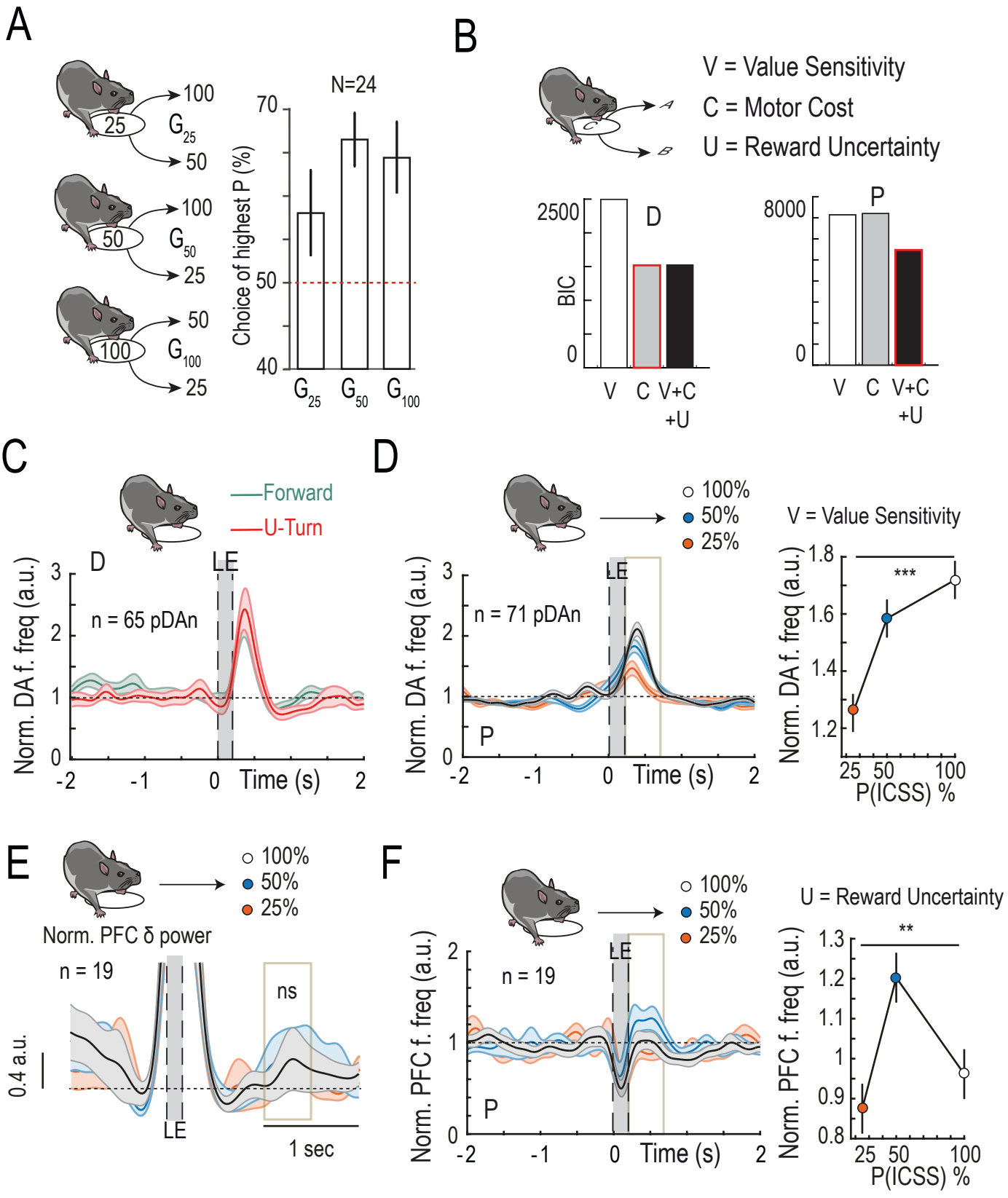

A
25
100
50
G25
50
100
25
G50
100
50
25
G100
N=24
Choice of highest P (%)
70
50
40
G25
G50
G100
B
V = Value Sensitivity
C = Motor Cost
U = Reward Uncertainty
2500
BIC
0
D
V
C
V+C
+U
8000
0
P
V
C
V+C
+U
C
Forward
U-Turn
D
LE
n = 65 pDAn
Norm. DA f. freq (a.u.)
3
2
1
0
-2
-1
0
Time (s)
2
D
100%
50%
25%
LE
n = 71 pDAn
P
Norm. DA f. freq (a.u.)
Time (s)
V = Value Sensitivity
***
1.8
1.6
1.4
1.2
Norm. DA f. freq (a.u.)
25
50
100
P(ICSS) %
E
100%
50%
25%
Norm. PFC δ power
n = 19
ns
0.4 a.u.
LE
1 sec
F
100%
50%
25%
LE
n = 19
P
Norm. PFC f. freq (a.u.)
2
1
0
-2
-1
0
Time (s)
2
U = Reward Uncertainty
**
1.3
1.2
1.1
1
0.9
0.8
Norm. PFC f. freq (a.u.)
25
50
100
P(ICSS) %

Figure 5

Figure 5: Complementary encoding of choices, value and cost by VTA, OFC and PFC. (A). Left: Scheme of the three gambles mice are facing in the probabilistic setting. Right: Proportion of choices of the location associated with the highest reward probability for each gamble (G_{25}: 100% vs 50%, G_{50}: 100% vs 25% and G_{100}: 50% vs 25%) at the end of the P setting (N=24). **(B).** Bayesian Information Criteria (BIC) computed using three models of choice selection at the end of the D setting and of the P setting (V: Softmax with β only, C: Softmax κ only, V+C+U: Softmax with β, κ and φ, See Methods). Red box surrounds the smaller BIC, indicating the best fit. **(C).** Normalized firing frequency (a.u.) of DA cells at the end of the D setting, centered on location entry. Trials are sorted according to the direction of the next choice compared to the previous one (U-turn indicating that mice performed a directional change by going back to the previous location, in contrast to forward). **(D).** Left: Normalized firing frequency (a.u.) of DA cells at the end of the P setting, centered on location entry. Trials are sorted according to the chosen location. Grey box indicates the quantification window. Right: Quantification of the DA firing frequency according to the reward probability of the goal. **(E).** Same as (D) left for normalized δ (3-6 Hz) power (a.u.) of PFC. **(F).** Same as (D) for normalized population firing frequency (a.u.) in the PFC.

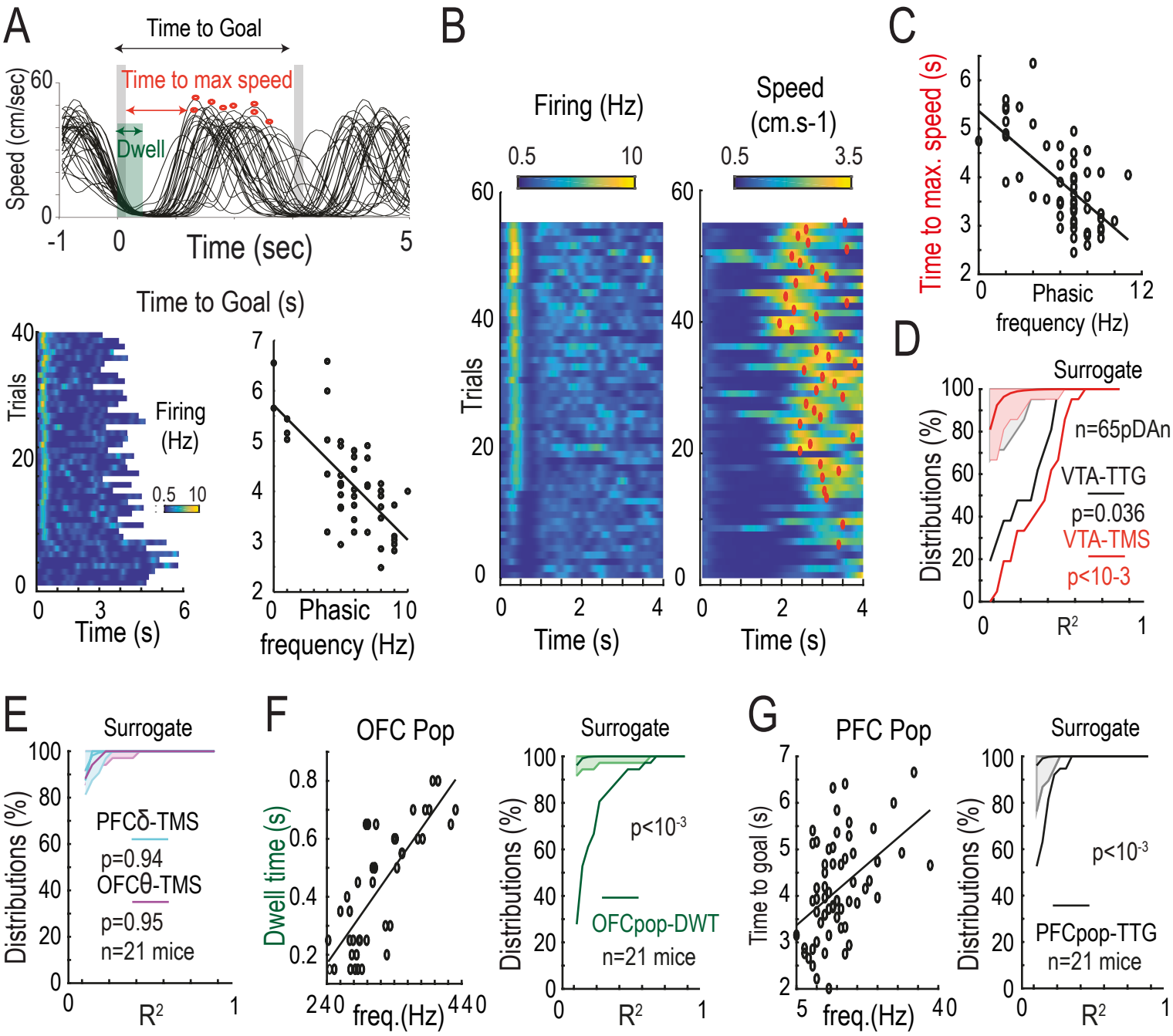

A
Time to Goal
Time to max speed
Dwell
Speed (cm/sec)
Time (sec)
Time to Goal (s)
Trials
Firing (Hz)
Time (s)
Phasic frequency (Hz)
B
Firing (Hz)
Speed (cm.s-1)
Trials
Time (s)
C
Time to max. speed (s)
Phasic frequency (Hz)
D
Surrogate
n=65pDAn
VTA-TTG
p=0.036
VTA-TMS
p<10-3
Distributions (%)
E
Surrogate
PFCδ-TMS
p=0.94
OFCθ-TMS
p=0.95
n=21 mice
F
OFC Pop
Dwell time (s)
freq.(Hz)
Surrogate
p<10-3
OFCpop-DWT
n=21 mice
G
PFC Pop
Time to goal (s)
freq.(Hz)
Surrogate
p<10-3
PFCpop-TTG
n=21 mice

Figure 6

Figure 6: Correlates of self-initiation, invigoration and pacing in OFC VTA and PFC respectively. (A). Top: Example of instantaneous speed profile for one mouse in the D setting. Grey boxes indicate ICSS durations, green boxes dwell times, red dots the maximal speed within trials. Bottom: Example of firing frequency for a pDA cell in the D setting with trials sorted from the smallest to the highest bursting frequency after trial initiation (left), and relation between the time to goal and the frequency of the early bursting activity (right). **(B).** Example of firing frequency for another pDA cell in the D setting with trials sorted from the smallest to the highest bursting frequency after trial initiation (left), and the instantaneous speed profile associated for each trial (right), red dots indicating the maximal speed within trials. **(C).** Relation between the time to maximal speed within trials and the phasic frequency for the cell showed in (B). **(D).** Distribution of correlation coefficients (R^2) for correlations between phasic frequency (Hz) and time to goal (black) and time to maximal speed (red) for n=65 cells. Surrogate data are generated by computing correlation with firing frequency taken at random timings within trials. **(E).** Distribution of correlation coefficients (R^2) for correlations between PFC δ power or OFC θ power and time to maximal speed at the end of the D setting. Surrogate data are generated by computing correlation with δ and θ powers taken at random timings within trials. **(F).** Left: Example of relation between the dwell time and the population firing frequency of OFC of one mouse at the end of the D setting, each dot representing a trial. Right: Distribution of correlation coefficients (R^2) for correlations between OFC population firing frequency and dwell time. Surrogate data are generated by computing correlation with firing frequency taken at random timings within trials. **(G).** Same as (F) for PFC population firing frequency and time to goal.

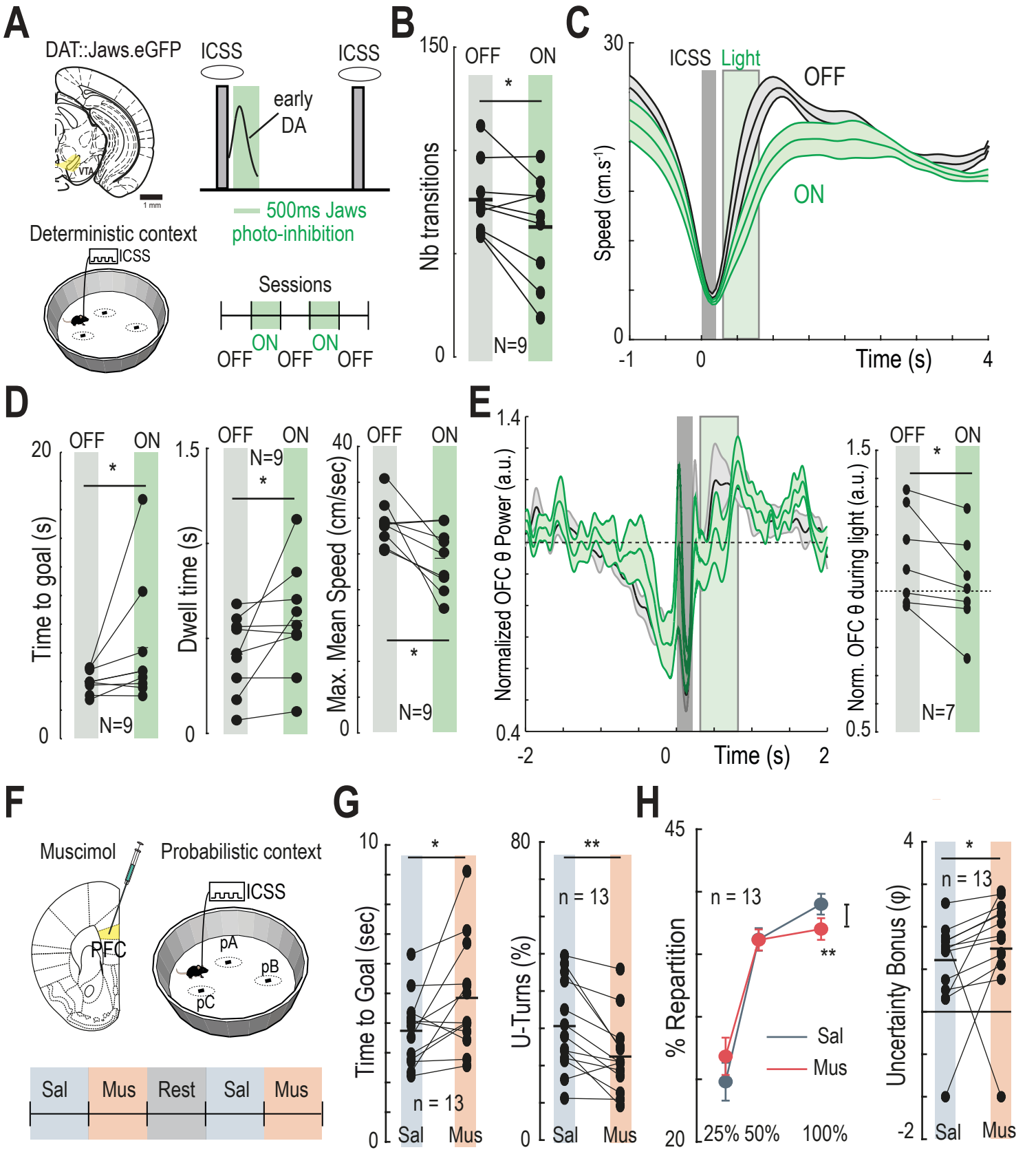

A
DAT::Jaws.eGFP
VTA
1 mm
ICSS
ICSS
early
DA
500ms Jaws
photo-inhibition
Deterministic context
ICSS
Sessions
ON
ON
OFF
OFF
OFF
B
150
0
OFF
ON
*
Nb transitions
N=9
C
30
0
ICSS
Light
OFF
ON
Speed (cm.s-1)
-1
0
Time (s)
4
D
20
0
OFF
ON
*
Time to goal (s)
N=9
1.5
0
OFF
ON
N=9
*
Dwell time (s)
40
0
OFF
ON
*
N=9
Max. Mean Speed (cm/sec)
E
1.4
0.4
Normalized OFC θ Power (a.u.)
-2
0
Time (s)
2
1.5
0.5
OFF
ON
*
N=7
Norm. OFC θ during light (a.u.)
F
Muscimol
Probabilistic context
ICSS
PFC
pA
pB
pC
Sal
Mus
Rest
Sal
Mus
G
10
0
*
Time to Goal (sec)
n = 13
Sal
Mus
80
0
**
U-Turns (%)
n = 13
Sal
Mus
H
45
20
n = 13
**
% Repartition
Sal
Mus
25%
50%
100%
4
-2
*
n = 13
Uncertainty Bonus (φ)
Sal
Mus

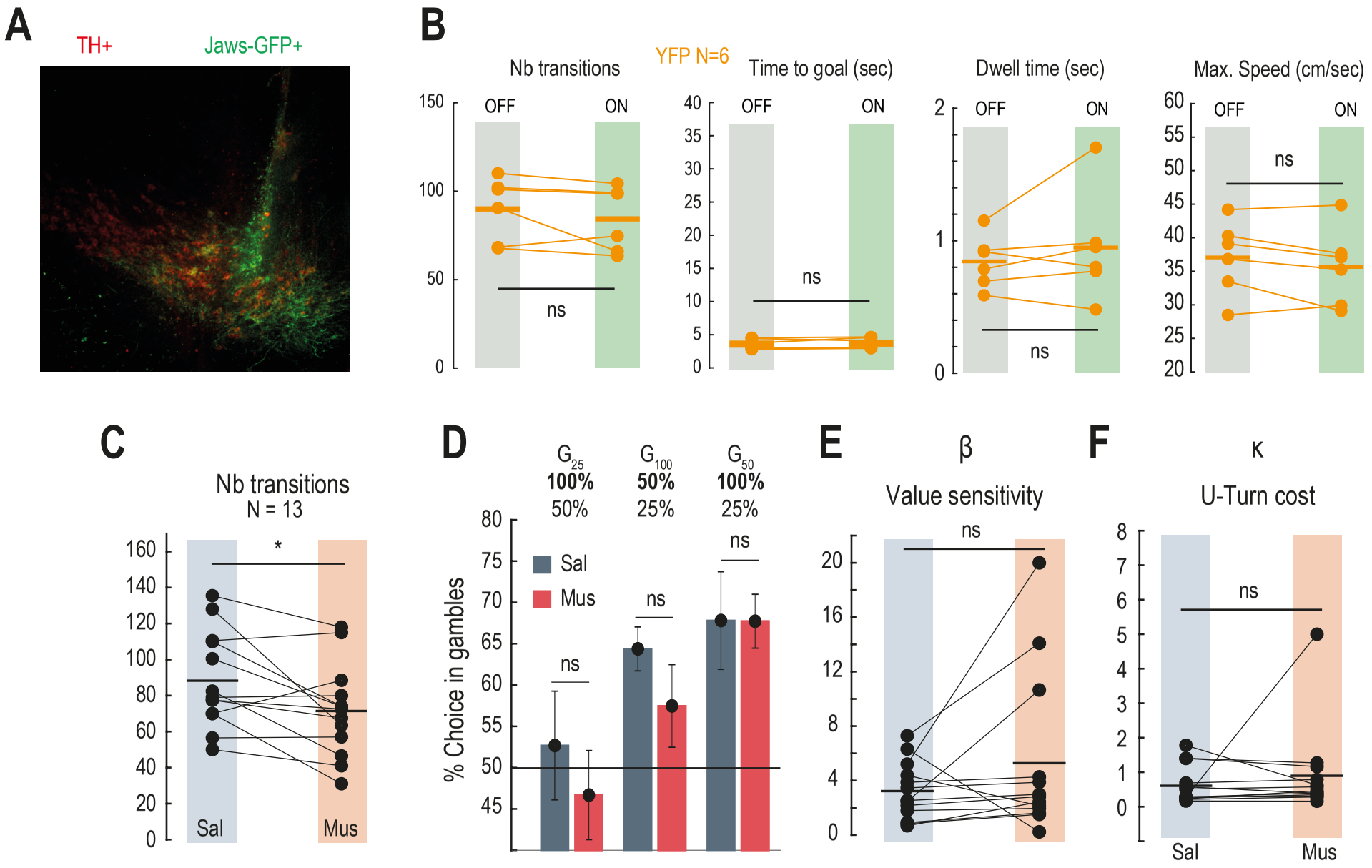
Figure 7

Figure 7: Causal manipulations. (A). Schematic representation of the manipulation experiment using the inhibitory opsin Jaws. Jaws is expressed unilaterally in the VTA of DATiCre mice using CreLox strategy. The light was applied continuously for 500 ms, after the ICSS (see Methods), in order to suppress the phasic DA activity. Mice underwent at the end of the D setting, a succession of OFF (light OFF) and ON (light ON) sessions. **(B).** Effect of OFF or ON light stimulations on the number of transitions (*i.e.* of rewards) obtained. (paired t-test, stat). **(C).** Effect of OFF or ON light stimulations on instantaneous speed profile within trials. Grey box indicates the ICSS duration and green box the light duration. **(D).** Effect of OFF or ON light stimulations on the time to goal (left), the dwell time (middle) and the maximal speed (right). **(E).** Left: Effect of OFF or ON light stimulations on the normalized OFC θ power (7-14 Hz). Grey box indicates the ICSS duration and green box the light duration. Right: Quantification of OFC θ power during light duration. **(F).** Schematic representation of the PFC inactivation experiment using bilateral muscimol infusion at the end of the P setting. Mice underwent a succession of sessions following saline or muscimol infusions. **(G).** Effect of muscimol on the time to goal (left) and the proportion of U-turns (right). **(H).** Effect of muscimol on the choice repartition between the three locations (left) and the fitted φ parameter (uncertainty bonus) obtained using the Softmax based on three parameters β, φ and κ (see Methods).

Supplementary information

Self-generation of goal-directed choices in a distributed dopaminergic and prefrontal circuit

Bousseyrol E #, Didienne S #, Takillah S, Prévost-Solié C, Come M, Ahmed Yahia T, Tricoire L, Dalkara D, Mourot A, Naudé J* & Faure Ph*



Supplementary Figure 1

Supplementary Figure 1:

(A). Image of TH+ and Jaws-GFP+ expressing cells in the VTA of a DATiCre mouse after expression of the inhibitory Jaws opsin. **(B).** Effect of the OFF-ON light application on (from left to right) the number of transitions, the time to goal, the dwell time and the maximum speed, for control YFP mice, at the end of the D setting. **(C).** Effect of the muscimol infusion on the number of transitions at the end of the P setting. **(D).** Effect of the muscimol infusion on the proportion of exploitation choice in the three gambles G_{25}, G_{100} and G_{50}. **(E).** Effect of muscimol infusion on the fitted β (value sensitivity) (left) and κ (U-turn cost) (right) parameters obtained using the Softmax based on three parameters β, φ and κ (see Methods).

References

Amarante, L.M., and Laubach, M. (2021). Coherent theta activity in the medial and orbital frontal cortices encodes reward value. Elife 10, e63372.

Anselme, P., Robinson, M.J.F., and Berridge, K.C. (2013). Reward uncertainty enhances incentive salience attribution as sign-tracking. Behav Brain Res 238, 53–61.

Bach, D.R., and Dolan, R.J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. 1–15.

Barter, J.W., Li, S., Lu, D., Bartholomew, R.A., Rossi, M.A., Shoemaker, C.T., Salas-Meza, D., Gaidis, E., and Yin, H.H. (2015). Beyond reward prediction errors: the role of dopamine in movement kinematics. Frontiers in Integrative Neuroscience 9, 1–22.

Belkaid, M., Bousseyrol, E., Cuttoli, R.D., Dongelmans, M., Duranté, E.K., Yahia, T.A., Didienne, S., Hanesse, B., Come, M., Mourot, A., et al. (2020). Mice adaptively generate choice variability in a deterministic task. Communications Biology 3, 1–9.

Berke, J.D. (2018). What does dopamine mean? 21, 787–793.

Cai, X., and Padoa-Schioppa, C. (2014). Contributions of Orbitofrontal and Lateral Prefrontal Cortices to Economic Choice and the Good-to-Action Transformation. Neuron 81, 1140–1151.

Carlezon, W.A., and Chartoff, E.H. (2007). Intracranial self-stimulation (ICSS) in rodents to study the neurobiology of motivation. Nature Protocols 2, 2987–2995.

Cavanagh, J.F., and Frank, M.J. (2014). Frontal theta as a mechanism for cognitive control. Trends Cogn Sci 18, 414–421.

Chuong, A.S., Miri, M.L., Busskamp, V., Matthews, G.A.C., Acker, L.C., Sørensen, A.T., Young, A., Klapoetke, N.C., Henninger, M.A., Kodandaramaiah, S.B., et al. (2014). Noninvasive optical inhibition with a red-shifted microbial rhodopsin.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. Philosophical Transactions Royal Soc B Biological Sci 362, 1585–1599.

Cisek, P. (2012). Making decisions through a distributed consensus. Current Opinion in Neurobiology 22, 927–936.

Coddington, L.T., and Dudman, J.T. (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci 21, 1563–1573.

Cohen, J.D., McClure, S.M., and Yu, A.J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences 362, 933–942.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. Nature 577, 671–675.

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

Dayan, P., and Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. 8, 429–453.

Dayan, P., and Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. Current Opinion in Neurobiology 18, 185–196.

Dongelmans, M., Cuttoli, R.D., Nguyen, C., Come, M., Duranté, E.K., Lemoine, D., Britto, R., Yahia, T.A., Mondoloni, S., Didienne, S., et al. (2021). Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in a foraging task. Biorxiv 2021.01.29.428835.

Ellwood, I.T., Patel, T., Wadia, V., Lee, A.T., Liptak, A.T., Bender, K.J., and Sohal, V.S. (2017). Tonic or Phasic Stimulation of Dopaminergic Projections to Prefrontal Cortex Causes Mice to Maintain or Deviate from Previously Learned Behavioral Strategies. J Neurosci 37, 8315–8329.

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. Nature Neuroscience 19, 479–486.

Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science (New York, N.Y.) 299, 1898–1902.

Fischbach-Weiss, S., Reese, R.M., and Janak, P.H. (2018). Inhibiting Mesolimbic Dopamine Neurons Reduces the Initiation and Maintenance of Instrumental Responding. Neuroscience 372, 306–315.

Fobbs, W.C., and Mizumori, S.J.Y. (2014). Cost-benefit decision circuitry: proposed modulatory role for acetylcholine. Progress in Molecular Biology and Translational Science 122, 233–261.

Fujisawa, S., and Buzsáki, G. (2011). A 4 Hz Oscillation Adaptively Synchronizes Prefrontal, VTA, and Hippocampal Activities. Neuron 72, 153–165.

Gershman, S.J., and Uchida, N. (2019). Believing in dopamine. Nature Reviews Neuroscience 40, 373.

Glimcher, P.W. (2011). Quantification of Behavior Sackler Colloquium: Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences of the United States of America.

Haber, S.N., and Knutson, B. (2010). The Reward Circuit: Linking Primate Anatomy and Human Imaging. Neuropsychopharmacol 35, 4–26.

Howe, M.W., and Dombeck, D.A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. Nature 535, 505–510.

Hunt, L.T., and Hayden, B.Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. Nat Rev Neurosci 18, 172–182.

Jin, X., and Costa, R.M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. Nature 466, 457–462.

Jo, Y.S., and Mizumori, S.J.Y. (2015). Prefrontal Regulation of Neuronal Activity in the Ventral Tegmental Area. Cereb Cortex 26, 4057–4068.

Jones, J.L., Esber, G.R., McDannald, M.A., Gruber, A.J., Hernandez, A., Mirenzi, A., and Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science (New York, N.Y.) 338, 953–956.

Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature 455, 227–231.

Klaus, A., Silva, J.A. da, and Costa, R.M. (2016). What, If, and When to Move: Basal Ganglia Circuits and Self-Paced Action Initiation. Annu Rev Neurosci 42, 1–25.

Knyazev, G.G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. Neurosci Biobehav Rev 36, 677–695.

Kringelbach, M.L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. Nature Reviews Neuroscience 6, 691–702.

Lak, A., Okun, M., Moss, M.M., Gurnani, H., Farrell, K., Wells, M.J., Reddy, C.B., Kepecs, A., Harris, K.D., and Carandini, M. (2020). Dopaminergic and Prefrontal Basis of Learning from Sensory Confidence and Reward Value. Neuron 105, 700-711.e6.

Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. Neuron 64, 431–439.

Marder, E., and Bucher, D. (2007). Understanding Circuit Dynamics Using the Stomatogastric Nervous System of Lobsters and Crabs. Annual Review of Physiology 69, 291–316.

Miller, E.K., and Wallis, J.D. (2013). Fundamental Neuroscience (Fourth Edition). Vii Behav Cognitive Neurosci 1069–1089.

Naudé, J., Tolu, S., Dongelmans, M., Torquet, N., Valverde, S., Rodriguez, G., Pons, S., Maskos, U., Mourot, A., Marti, F., et al. (2016). Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nature Neuroscience 19, 471–478.

O'Neill, M., and Schultz, W. (2010). Coding of Reward Risk by Orbitofrontal Neurons Is Mostly Distinct from Coding of Reward Value. Neuron 68, 789–800.

Onge, J.R.S., and Floresco, S.B. (2010). Prefrontal cortical contribution to risk-based decision making. Cerebral Cortex (New York, NY : 1991) 20, 1816–1828.

Padoa-Schioppa, C. (2011). Neurobiology of Economic Choice: A Good-Based Model. Annual Review of Neuroscience 34, 333–359.

Padoa-Schioppa, C., and Conen, K.E. (2017). Orbitofrontal Cortex: A Neural Circuit for Economic Decisions. Neuron 96, 736–754.

Park, J., and Moghaddam, B. (2017). Risk of punishment influences discrete and coordinated encoding of reward-guided actions by prefrontal cortex and VTA neurons. Elife 6, e30056.

Passingham, R.E., Bengtsson, S.L., and Lau, H.C. (2010). Medial frontal cortex: from self-generated action to reflection on one's own performance. Trends Cogn Sci 14, 16–21.

Pezzulo, G., and Cisek, P. (2016). Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. Trends Cogn Sci 20, 414–424.

Puig, M.V., and Miller, E.K. (2012). The Role of Prefrontal Dopamine D1 Receptors in the Neural Mechanisms of Associative Learning. Neuron 74, 874–886.

Puig, M.V., Antzoulatos, E.G., and Miller, E.K. (2014). Prefrontal dopamine in associative learning and memory. Neuroscience 282C, 217–229.

Quyen, M.L.V., and Bragin, A. (2007). Analysis of dynamic brain oscillations: methodological advances. Trends Neurosci 30, 365–373.

Rescorla, R.A., and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. pp. 64–99.

Roesch, M.R., Calu, D.J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience 10, 1615–1624.

Rushworth, M.F.S., and Behrens, T.E.J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. Nature Neuroscience 11, 389–397.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70, 1054–1069.

Salamone, J.D., and Correa, M. (2012). The Mysterious Motivational Functions of Mesolimbic Dopamine. Neuron 76, 470–485.

Schuck-Paim, C., Pompilio, L., and Kacelnik, A. (2004). State-Dependent Decisions Cause Apparent Violations of Rationality in Animal Choice. PLoS biology 2, e402.

Schultz, W. (2007). Multiple dopamine functions at different time courses. Annual Review of Neuroscience 30, 259–288.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. Science 275, 1593–1599.

Sharpe, M.J., Stalnaker, T., Schuck, N.W., Killcross, S., Schoenbaum, G., and Niv, Y. (2018). An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. Annu Rev Psychol 70, 1–24.

Silva, J.A. da, Tecuapetla, F., Paixão, V., and Costa, R.M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. Nature 554, 1–21.

Stalnaker, T.A., Cooch, N.K., and Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. Nat Neurosci 18, 620–627.

Starkweather, C.K., Gershman, S.J., and Uchida, N. (2018). The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. Neuron 98, 616-629.e6.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nature Neuroscience 1–10.

Steinmetz, N.A., Zatka-Haas, P., Carandini, M., and Harris, K.D. (2019). Distributed coding of choice, action and engagement across the mouse brain. Nature 576, 266–273.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning (MIT Press).

Syed, E., Grima, L.L., Magill, P.J., and Bogacz, R. (2015). Action initiation shapes mesolimbic dopamine encoding of future rewards. Nature.

Takahashi, Y.K., Roesch, M.R., Stalnaker, T.A., Haney, R.Z., Calu, D.J., Taylor, A.R., Burke, K.A., and Schoenbaum, G. (2009). The Orbitofrontal Cortex and Ventral Tegmental Area Are Necessary for Learning from Unexpected Outcomes. Neuron 62, 269–280.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., and Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat Neurosci 14, 1590–1597.

Takahashi, Y.K., Batchelor, H.M., Liu, B., Khanna, A., Morales, M., and Schoenbaum, G. (2017). Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. Neuron 95, 1395-1405.e3.

Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. Science (New York, N.Y.) 307, 1642–1645.

Tobler, P.N., Christopoulos, G.I., O'Doherty, J.P., Dolan, R.J., and Schultz, W. (2009). Risk-dependent reward value signal in human prefrontal cortex. Proc National Acad Sci 106, 7185–7190.

Verharen, J.P.H., Jong, J.W. de, Roelofs, T.J.M., Huffels, C.F.M., Zessen, R. van, Luijendijk, M.C.M., Hamelink, R., Willuhn, I., Ouden, H.E.M. den, Plasse, G. van der, et al. (2018). A neuronal mechanism underlying decision-making deficits during hyperdopaminergic states. Nat Commun 9, 731.

Wassum, K.M., Ostlund, S.B., and Maidment, N.T. (2012). Phasic Mesolimbic Dopamine Signaling Precedes and Predicts Performance of a Self-Initiated Action Sequence Task. Biol Psychiat 71, 846–854.

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron 81, 267–279.

Yoo, S.B.M., and Hayden, B.Y. (2018). Economic Choice as an Untangling of Options into Actions. Neuron 99, 434–447.

Yu, A.J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron 46, 681–692.

Yun, M., Kawai, T., Nejime, M., Yamada, H., and Matsumoto, M. (2020). Signal dynamics of midbrain dopamine neurons during economic decision-making in monkeys. Sci Adv 6, eaba4962.

Chapitre 6 : Deuxième publication

Mice adaptively generate choice variability in a deterministic task. Belkaid M, Bousseyrol E, Durand-de Cuttoli R, Dongelmans M, Duranté E K, Ahmed Yahia T, Didienne S, Hannesse B, Come M, Mourot A, Naudé J, Sigaud O* & Faure P*. Communications Biology. 3, 34 (2020).

Comme exposé en introduction et dans le premier article de thèse, le cerveau génère des prédictions sur son environnement à partir des informations acquises par l'expérience pour générer des choix comportementaux. Bon nombre de nos choix consistent ainsi à répéter des actions réussies dans le passé, et pour lesquelles le circuit de la récompense a créé une association positive : il s'agit là d'un apprentissage par renforcement (RL) classique. Quand un animal répète des choix précédemment renforcés, on parle d'exploitation. Cependant, dans certaines situations nouvelles, incertaines ou complexes, mais aussi lorsque l'on est confronté à une même situation de manière répétée, produire un comportement en apparence sous-optimal, inhabituel, variable ou imprévisible peut constituer un avantage stratégique. Ces comportements fréquents et spontanés témoignent d'une variabilité intra-individuelle et peuvent être qualifiés d'exploration, même si le gain d'information n'est pas toujours évident.

Dans le cadre de la prise de décision économique, on définit souvent l'exploration comme l'option alternative à l'exploitation. Mais l'exploration et la variabilité sont des concepts plus vastes qui s'observent aussi dans des comportements naturels : on parle par exemple de curiosité, d'investigation, de divertissement, de récréation ou de jeu. C'est aussi le cas des trajectoires des proies pour échapper à un prédateur ou d'humains qui jouent à des jeux de compétition. La génération volontaire de comportements aléatoires est réputée difficile, notamment chez l'humain. Pour certains, la variabilité comportementale trouve son origine dans du bruit (sensoriel ou moteur) ou dans des erreurs du système de décision. Il n'est donc toujours pas clair si le cerveau est capable de générer de la variabilité délibérément, intrinsèquement, et sans forcément que celle-ci soit dirigée, biaisée, vers des stimuli particuliers de l'environnement.

Dans ce second travail, nous nous sommes intéressés à une autre forme d'exploration, intrinsèque à l'animal et non déterminée par les contingences de l'environnement, aussi dite « exploration aléatoire ». Pour cela, nous avons donc développé au laboratoire une autre règle de distribution des récompenses dans la tâche 3 points (règle de complexité, cf. Chapitre 4) afin de favoriser la variabilité et l'exploration aléatoire dans des séquences de choix : les choix de l'animal ne sont récompensés que s'ils sont imprévisibles. Les récompenses sont donc délivrées suivant une règle déterministe, il n'y a pas de hasard dans la tâche. La variabilité du comportement ne peut donc pas s'expliquer par une variabilité dans l'environnement.

Pour maximiser leur gain, les souris ont deux options : soit utiliser leur mémoire (se rappeler d'une séquence de N choix efficaces) pour battre la règle et tendre vers 100% de succès mais à un coût computationnel très élevé, soit choisir au hasard ce qui permet de performer assez bien et obtenir 75% de succès pour un coût cognitif faible. Dans ce travail, nous avons constaté que les souris augmentaient progressivement la variabilité de leurs séquences de choix. Plus important encore, les résultats expérimentaux et la modélisation suggèrent qu'elles atteignent la meilleure performance possible en utilisant une sélection aléatoire, sans utiliser leur mémoire. Ces résultats démontrent qu'il existe un processus de décision sous-tendant la génération de variabilité et d'aléatoire de manière auto-générée et indépendante des règles de l'environnement.

COMMUNICATIONS BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-020-0759-x OPEN

Corrected: Author Correction

Mice adaptively generate choice variability in a deterministic task

Marwen Belkaid[1], Elise Bousseyrol[2], Romain Durand-de Cuttoli[2], Malou Dongelmans[2], Etienne K. Duranté[2], Tarek Ahmed Yahia[2], Steve Didienne[2], Bernadette Hanesse[2], Maxime Come[2], Alexandre Mourot[2], Jérémie Naudé[2], Olivier Sigaud[1,3] & Philippe Faure[2,3*]

Can decisions be made solely by chance? Can variability be intrinsic to the decision-maker or is it inherited from environmental conditions? To investigate these questions, we designed a deterministic setting in which mice are rewarded for non-repetitive choice sequences, and modeled the experiment using reinforcement learning. We found that mice progressively increased their choice variability. Although an optimal strategy based on sequences learning was theoretically possible and would be more rewarding, animals used a pseudo-random selection which ensures high success rate. This was not the case if the animal is exposed to a uniform probabilistic reward delivery. We also show that mice were blind to changes in the temporal structure of reward delivery once they learned to choose at random. Overall, our results demonstrate that a decision-making process can self-generate variability and randomness, even when the rules governing reward delivery are neither stochastic nor volatile.

[1] Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique (ISIR), 75005 Paris, France. [2] Sorbonne Université, INSERM, CNRS, Neuroscience Paris Seine - Institut de Biologie Paris Seine (NPS - IBPS), 75005 Paris, France. [3] These authors contributed equally: Olivier Sigaud, Philippe Faure. *email: phfaure@gmail.com

Principles governing random behaviors are still poorly understood, despite well-known ecological examples ranging from vocal and motor babbling in trial-and-error learning[1,2] to unpredictable behavior in competitive setups (e.g preys-versus-predators[3] or humans competitive games[4]). Dominant theories of behavior and notably reinforcement learning (RL) rely on exploitation, namely the act of repeating previously rewarded actions[5,6]. In this context, choice variability is associated with exploration of environmental contingencies. Directed exploration aims at gathering information about environmental contingencies[7,8], whereas random exploration introduces variability regardless of the contingencies[9,10]. Studies have shown that animals are able to produce variable, unpredictable choices[11,12], especially when the reward delivery rule changes[13–15], is stochastic[9,16,17] or is based on predictions about their decisions[18,19]. However, even approaches based on the prediction of the animal behavior[18,19] keep the possibility to distribute reward stochastically—for example if no systematic bias in the animal's choice behavior has been found[19,20]. Thus, because of the systematic use of volatile or probabilistic contingencies, it has remained difficult to experimentally isolate variability generation from environmental conditions. To test the hypothesis that animals can adaptively adjust the randomness of their behavior, we implemented a task where the reward delivery rule is deterministic, predetermined and identical for all animals, but where a purely random choices strategy is successful.

Results

Mice can generate variable decisions in a complex task. Mice were trained to perform a sequence of binary choices in an open-field where three target locations were explicitly associated with rewards delivered through intra-cranial self-stimulation (ICSS) in the medial forebrain bundle. Importantly, mice could not receive two consecutive ICSS at the same location. Thus, they had to perform a sequence of choices[16] and at each location to choose the next target amongst the two remaining alternatives (Fig. 1a). In the training phase, all targets had a 100% probability of reward. We observed that after learning, mice alternated between rewarding locations following a stereotypical circular scheme interspersed with occasional changes in direction, referred to as U-turn (Fig. 1b). Once learning was stabilized, we switched to the complexity condition, in which reward delivery was non-stochastic and depended on sequence variability. More precisely, we calculated the Lempel-Ziv (LZ) complexity[21] of choice subsequences of size 10 (9 past choices + next choice) at each trial. Animals were rewarded when they chose the one target (out of the two options) associated with the highest complexity (given the previous nine choices). Despite its difficulty, this task is fully deterministic. Indeed, mice were asked to move along a tree of binary choices (see Fig. 1a) where some paths ensured 100% rewards. Whether each node was rewarded or not was predetermined in advance. Thus, choice variability could not be imputed to the inherent stochasticity of the outcomes. For each trial, if choosing randomly, the animal had either 100% or 50% chance of being rewarded depending on whether the two subsequences of size 10 (=9 past choices + 1 choice out of 2 options) had equal or unequal complexities. Another way to describe the task is thus to consider all possible situations, not as sequential decisions made by the animal during the task but as the set of all possible subsequences of size 10 of which the algorithm may evaluate the complexity. From this perspective, there is an overall 75% probability of being rewarded if subsequences are sampled uniformly (Fig. 1a). To summarize, theoretically, while a correct estimation of the complexity of the sequence leads to a success rate of 100%, a pure random selection at each step leads to 75% of success, and a repetitive sequence (e.g. A,B,C,A,B,C,…) grants no reward.

Unlike the stereotypical circular scheme observed during training, at the end of the complexity condition, choice sequence became more variable (Fig. 1b). We found that mice progressively increased the variability of their choice sequences and thus their success rate along sessions (Fig. 1c). This increased variability in the generated sequences was demonstrated by an increase in the normalized LZ-complexity measure (hereafter NLZcomp) of the session sequences, a decrease in an entropy measure based on recurrence plot quantification and an increase in the percentage of U-turns (Fig. 1d). Furthermore, in the last session, 65.5% of the sequences were not significantly different from surrogate sequences generated randomly (Supplementary Fig 1a). The success rate was correlated with the NLZcomp of the entire session of choice sequences (Fig. 1e), suggesting that mice increased their reward through an increased variability in their choice. The increase in success rate was associated with an increase of the percentage of U-turns (Fig. 1d), yet mice performed a suboptimal U-turn rate of 30%, below the 50% U-turn rate ensuring 100% of rewards (Supplementary Fig 1b).

Computational modeling indicates the use of random strategy. From a behavioral point of view, mice thus managed to increase their success rate in a highly demanding task. They did not achieve 100% success but reached performances that indicate a substantial level of variability. Given that the task is fully deterministic, the most efficient strategy would be to learn and repeat one (or a subset) of the 10-choice long sequences that are always rewarded. This strategy ensures the highest success rate but incurs a tremendous memory cost. On the other hand, a purely random selection is another appealing strategy since it is less costly and leads to about 75% of reward. To differentiate between the two strategies and better understand the computational principles underlying variability generation in mice, we examined the ability of a classical RL algorithm to account for the mouse decision-making process under these conditions.

As in classical reinforcement learning, state-action values were learned using the Rescorla-Wagner rule[22] and action selection was based on a softmax policy[5] (Fig. 2a; see "Methods"). Two adaptations were applied: (i) rewards were discounted by a U-turn cost κ in the utility function in order to reproduce mouse circular trajectories in the training phase; (ii) states were represented as vectors in order to simulate mouse memory of previous choices. By defining states as vectors including the history of previous locations instead of the current location alone, we were able to vary the memory size of simulated mice and to obtain different solutions from the model accordingly. We found that, with no memory (i.e. state = current location), the model learned equal values for both targets in almost all states (Fig. 2b). In contrast, and in agreement with classical RL, with the history of the nine last choices stored in memory, the model favored the rewarded target in half of the situations by learning higher values (approximately 90 vs 10%) associated with rewarded sequences of choices (Fig. 2b). This indicates that classical RL can find the optimal solution of the task if using a large memory. Furthermore, choosing randomly was dependent not only on the values associated with current choices, but also on the softmax temperature and the U-turn cost. The ratio between these two hyperparameter controls the level of randomness in action selection (see "Methods"). Intuitively, a high level of randomness leads to high choice variability and sequence complexity. But interestingly, the randomness hyperparameter had opposite effects on the model behavior with small and large memory sizes. While increasing the temperature always increased the

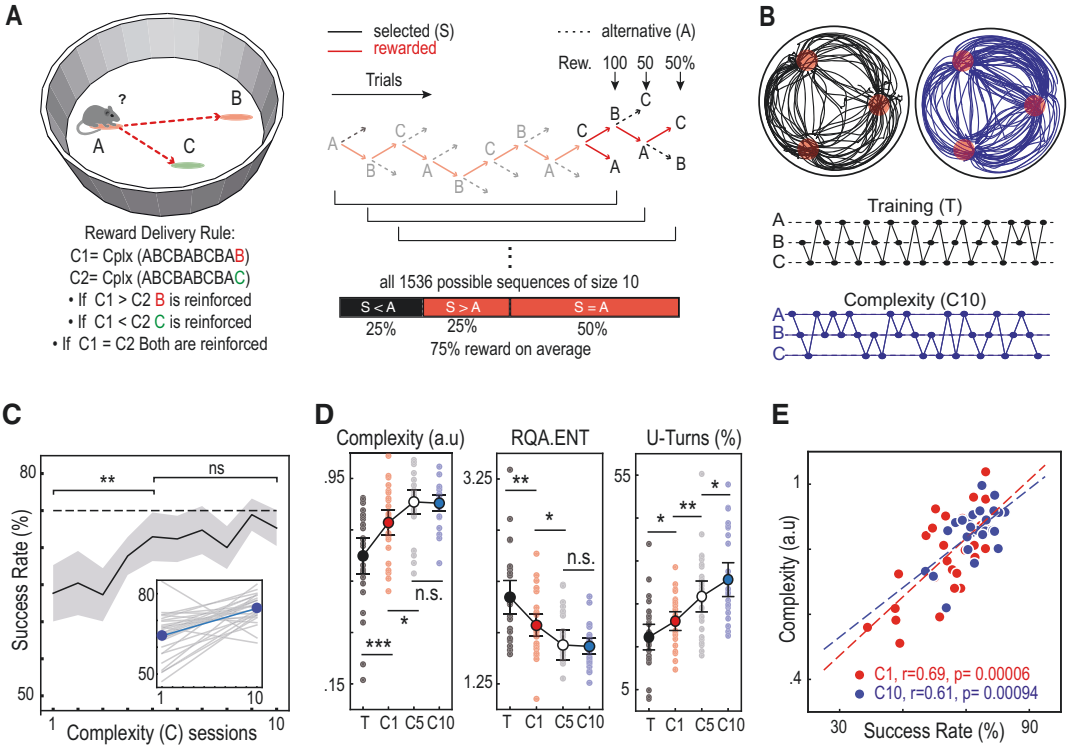


Fig. 1 Mice generate unpredictable decisions. a Left: task setting and complexity algorithm for reward delivery (see text) Right: tree structure of the task and reward distribution. **b** Typical trajectories in training (T) and complexity (C) conditions. **c** Increase of the success rate over sessions in the complexity setting. Mice improved their performance in the first sessions (c01 versus c05, $T = 223.5$, $p = 0.015$, Wilcoxon test) then reached a plateau (c05 versus c10, $t(25) = -0.43$, $p = 0.670$, paired t-test) close to the theoretical 75% success rate of random selection (c10, $t(25) = -1.87$, $p = 0.073$, single sample t-test). The shaded area represents a 95% confidence interval. Inset, linear regressions of individual mice performance increase for individual mice (gray line) and average progress (blue line). **d** Increase of the behavior complexity over sessions: the *NLZcomp* measure of complexity increased in the beginning (training versus c01, $T = 52$, $p = 0.0009$, Wilcoxon test, c01 versus c05, $t(26) = -2.67$, $p = 0.012$, paired t-test) before reaching a plateau (c05 versus c10, $T = 171$, $p = 0.909$, Wilcoxon test). The average complexity reached by animal is lower than 1 (c10, $t(25) = -9.34$, $p = 10\text{-}9$, single sample t-test), which corresponds to the complexity of random sequences. The RQA *ENT* entropy-based measure of complexity decreased over sessions (training versus c01, $t(26) = 2.81$, $p = 0.009$, paired t-test, c01 versus c05, $T = 92$, $p = 0.019$, Wilcoxon test, c05 versus c10, $T = 116$, $p = 0.13$, Wilcoxon test). The rate of U-turns increased over sessions (training versus c01, $t(26) = -2.21$, p = 0.036, c01 versus c05, $t(26) = -3.07$, $p = 0.004$, paired t-test, c05 versus c10, $T = 75$, $p = 0.010$, Wilcoxon test). Error bars represent 95% confidence intervals. **e** Correlation between individual success rate and complexity of mice sequences. Also noteworthy is the decrease in data dispersion in session c10 compared to c1. $N = 27$ in all sessions except c10 where $N = 26$.

complexity of choice sequences, it increased the success rate for small memory sizes but decreased it for larger memories (Fig. 2c). A boundary between the two regimes was found between memory sizes of 3 and 4.

Upon optimization of the model to fit mouse behavior, we found that their performance improvement over sessions was best accounted for by an increase of choice randomness using a small memory (Fig. 2d). This model captured mouse learning better than when using fixed parameters throughout sessions (Bayes factor = 3.46; see "Methods", and Supplementary Fig. 2a, b). The model with a memory of size 3 best reproduced mouse behavior (Fig. 2d), but only slightly better than versions with smaller memories (Supplementary Fig 2c). From a computational perspective, one possible explanation for the fact that although theoretically sufficient, a memory of size 1 fits less than size 3, is that state representation is overly simplified in the model. Accordingly, altering the model's state representation to make it more realistic should reduce the size of the memory needed to reproduce mice performances. To test this hypothesis, we used a variant of the model in which we manipulated state representation ambiguity: each of the locations *{A, B, C}* could be represented by $n \geq 1$ states, with $n = 1$ corresponding to unambiguous states (see "Methods"", and Fig. 2e). As expected, the model fitted better with a smaller memory as representation ambiguity was increased (Fig. 2e). We also found that the best fitting learning rate was higher with ambiguous representations while the randomness factor remained unchanged regardless of ambiguity level (Fig. 2e). This corroborates that the use of additional memory capacity by the model is due to the model's own limitations rather than an actual need to memorize previous choices. Hence, this computational analysis overall suggests that mice adapted the randomness parameter of their decision-making system to achieve more variability over sessions rather than remembered rewarded choice sequences. This conclusion was further reinforced by a series of behavioral arguments detailed below supporting the lack of memorization of choice history in their strategy.

Mice choose randomly without learning the task structure. We first looked for evidence of repeated choice patterns in mouse sequences using a Markov chain analysis (see "Methods"). We found that the behavior at the end of the complexity condition was Markovian (Fig. 3a). In other words, the information about the immediately preceding transition (i.e. to the left or to the right) was necessary to determine the following one (e.g. $p(L) \neq P(L|L)$) but looking two steps back was not informative on future decisions (e.g. $p(L|LL) \approx P(L|L)$) (see "Methods" Markov Chain

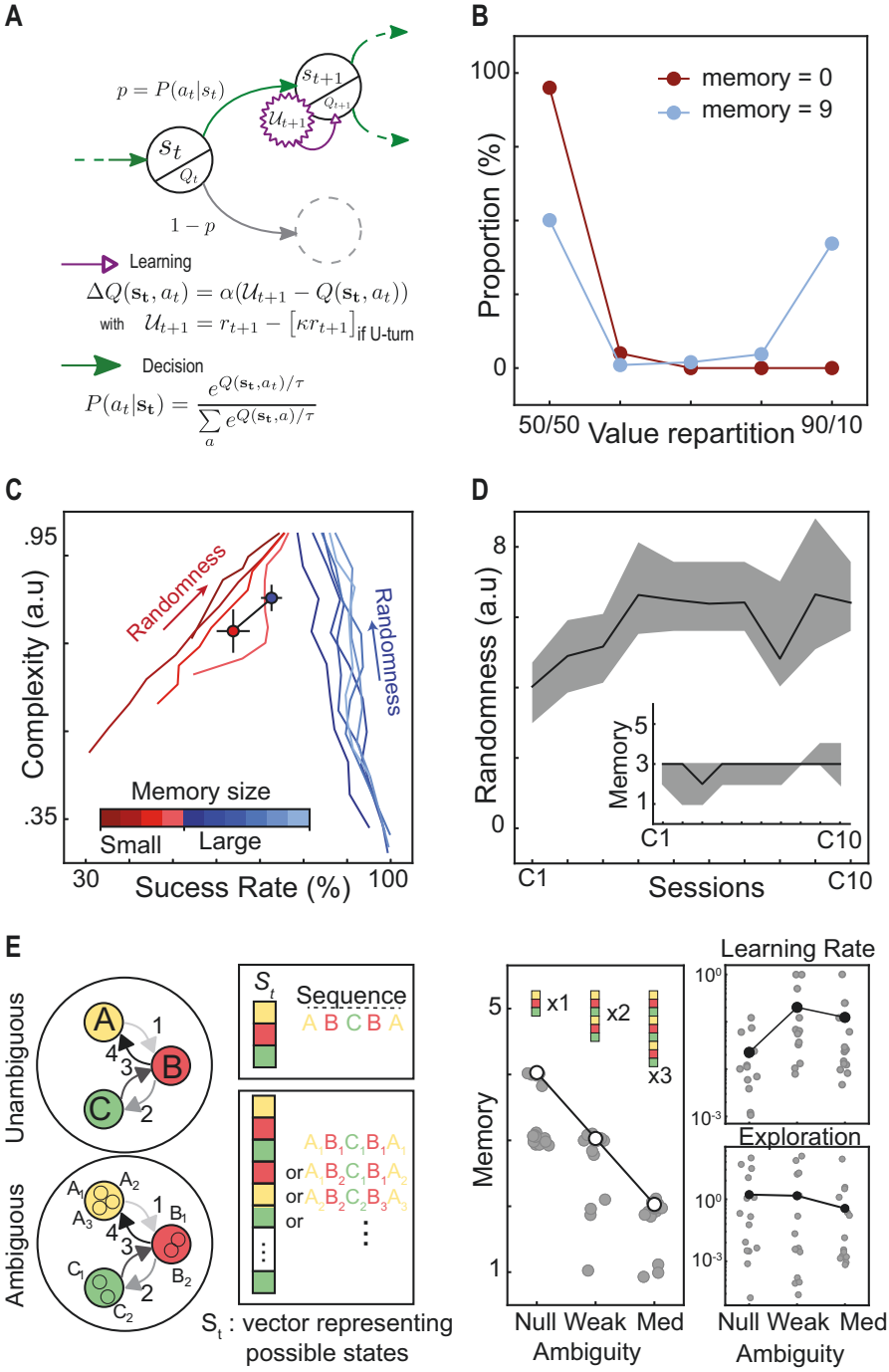


Fig. 2 Computational modeling suggests a memory-free pseudo-random selection behind mice choice variability. a Schematic illustration of the computational model fitted to mouse behavior. **b** Repartition of the values learned by the model with memory size equal to 0 or 9. **c** Influence of increased randomness on success rate and complexity for various memory sizes. Each line describes the trajectory followed by a model with a certain memory size (see color scale) when going from a low to high level of randomness (defined as τ/κ). Red and blue dots represent experimental data of mice in the last training and complexity sessions, respectively. **d** Model fitting results. With an increase of randomness and a small memory, the model fits the increase in mice performance. The shaded areas represent values of the 15 best parameter sets. Dark lines represent the average randomness value (continuous values) and the best fitting memory size (discrete values), respectively. **e** Schematic of ambiguous state representations and simulation results. The main simulations rely on an unambiguous representation of states in which each choice sequence is represented by one perfectly recognized code. With ambiguous states, the same sequence can be encoded by various representations. In the latter case, the model best fits mouse performance with a smaller memory (null, weak and medium ambiguity, $H = 27.21$, $p = 10\text{-}6$, Kruskal–Wallis test, null versus weak, $U = 136$, $p = 0.006$, weak versus med, $U = 139$, $p = 0.002$, Mann–Whitney test) and with a higher learning rate (null, weak, and medium ambiguity, $H = 7.61$, $p = 0.022$, Kruskal–Wallis test, null versus weak, $U = 45.5$, $p = 0.016$, null versus med, $U = 54$, $p = 0.026$, weak versus med, $U = 101$, $p = 0.63$, Mann–Whitney test) but a similar exploration rate (null, weak and medium ambiguity, $H = 3.64$, $p = 0.267$, Kruskal–Wallis test). Gray dots represent the 15 best fitting parameter sets. White dots represent the best fit in case of a discrete variable (memory) while black dots represent the average in case of continuous variables (temperature and learning rate). $N = 15$.

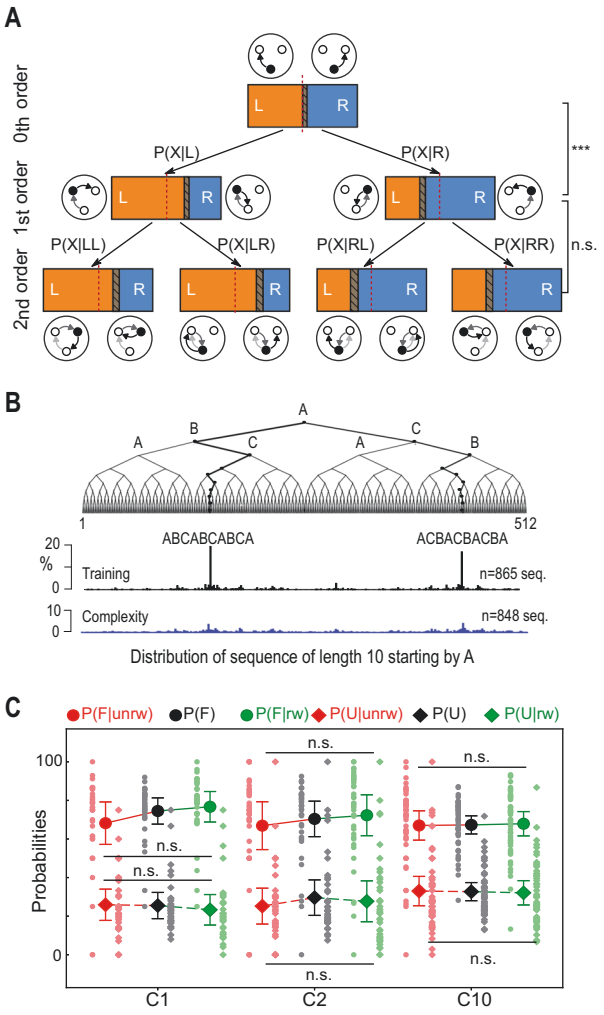


Fig. 3 Behavioral evidence of the absence of memorization in mouse choices. a Tree representation of the Markovian structure of mouse behavior in session c10 ($N = 26$). In the expression of probabilities, P(X) refers to P(L) (L Left) or P(R) (R right), whose repartition is illustrated in the horizontal bars (respectively in orange and blue). Dashed areas inside the bars represent overlapping 95% confidence intervals. The probability of a transition (i.e. to the left or to the right) is different from the probability of the same transition given the previous one ($p < 0.05$, paired t-test, see "Methods" for detailed analysis). However, the probability given two previous transitions is not different from the latter ($p > 0.05$, paired t-test, see "Methods" for detailed analysis). **b** Distribution of subsequences of length 10. **c** Absence of influence of rewards on mice decisions. P(F) and P(U) respectively refer to the probabilities of going forward (e.g. A → B → C) and making a U-turn (e.g. A → B → A). These probabilities were not different from the conditional probabilities given that the previous choice was rewarded or not ($p > 0.05$, Kruskal–Wallis test, see "Methods" for detailed analysis). This means that the change in mice behavior under the complexity condition was not stereotypically driven by the outcome of their choices (e.g. "u-turn if not rewarded"). Error bars in B represent 95% confidence intervals. $N = 34$ in c01, $N = 38$ in c02, and $N = 52$ in c10.

Analysis). The analysis of the distribution of subsequence of length 10 (see "Methods") provides an additional evidence of the lack of structure in the animals' choice sequence. Indeed, while at the end of the training, mice exhibit a peaky distribution with a strong preference for the highly repetitive circular patterns and their variants, the distribution was dramatically flattened under the complexity condition (Fig. 3b) demonstrating that mice behavior is much less structured in this setting. Furthermore, we tested whether mice use of a win-stay-lose-switch strategy[18]. Indeed, mice could have used this heuristic strategy when first confronted with the complexity condition after a training phase in which all targets were systematically rewarded. Changing directions in the absence of reward could have introduced enough variability in the animals' sequence to improve their success rate. Yet, we found that being rewarded (or not) had no effect on the next transition, neither at the beginning nor the end of the complexity condition (Fig. 3c; see "Methods"); thus eliminating another potential form of structure in mice behavior under the complexity rule.

To further support the notion that mice did not actually memorize rewarded sequences to solve the task, we finally performed a series of experiments to compare the animals' behavior under the complexity rule and under a probabilistic rule in which all targets were rewarded with a 75% probability (the same frequency reached at the end of the complexity condition). We first analyzed mice behavior when the complexity condition was followed by the probabilistic condition (Group 1 in Fig. 4a). We hypothesized that, if animals choose randomly at each node in the complexity setting (and thus do not memorize and repeat specific sequences), they would not detect the change of the reward distribution rule when switching to the probabilistic setting. In agreement with our assumption, we observed that as we switched to the probabilistic condition, animals did not modify their behavior although the optimal strategy would have been to avoid U-turns, as observed in the 100% reward setup used for training (Fig. 4b and Supplementary Fig 3a). Hence, after the complexity setting, mice were likely stuck in a "random" mode given that the global statistics of the reward delivery were conserved. In contrast, when mice were exposed to the probabilistic distribution of reward right after the training session (Group 2 in Fig. 4a), they slightly changed their behavior but mostly stayed in a circular pattern with few U-turns and low sequence complexity (Fig. 4b and Supplementary Fig 3a). Thus, animals from Group 2 exhibited lower sequence complexity and U-turn rate in the probabilistic condition than animals from Group 1, whether in the complexity or the probabilistic condition (Fig. 4c). The distribution of patterns of length 10 in the sequences performed by animals from Group 2 during the last probabilistic session shows a preference for repetitive circular patterns that is very similar to that observed at the end of the training; contrasting with the sequences performed by animals from Group 1 (Fig. 4d, e). A larger portion of sequences performed by animals from Group 1 were not different from surrogate sequences generated randomly in comparison with animals from Group 2 (Supplementary Fig 3b). Last, if the sequences performed by mice from Group 2 were executed under the complexity rule, these animals would have obtained lower success rate than animals from Group 1 in the complexity condition (Supplementary Fig. 3c).

In summary, mice behavior under the probabilistic condition changed markedly depending on the preceding condition and the strategy that the animal was adopting. This further supports our initial claim that stochastic experimental setups make it difficult to unravel the mechanisms underlying random behavior generation.

Discussion

The deterministic nature of complexity rule used in our experiments makes it possible to categorize animals' behavior into one of three possible strategies (i.e. repetitive, random or optimal based on sequence learning). This is crucial in understanding the underlying cognitive process leveraged by the animals. Importantly, this shall not be interpreted as implying that animals were

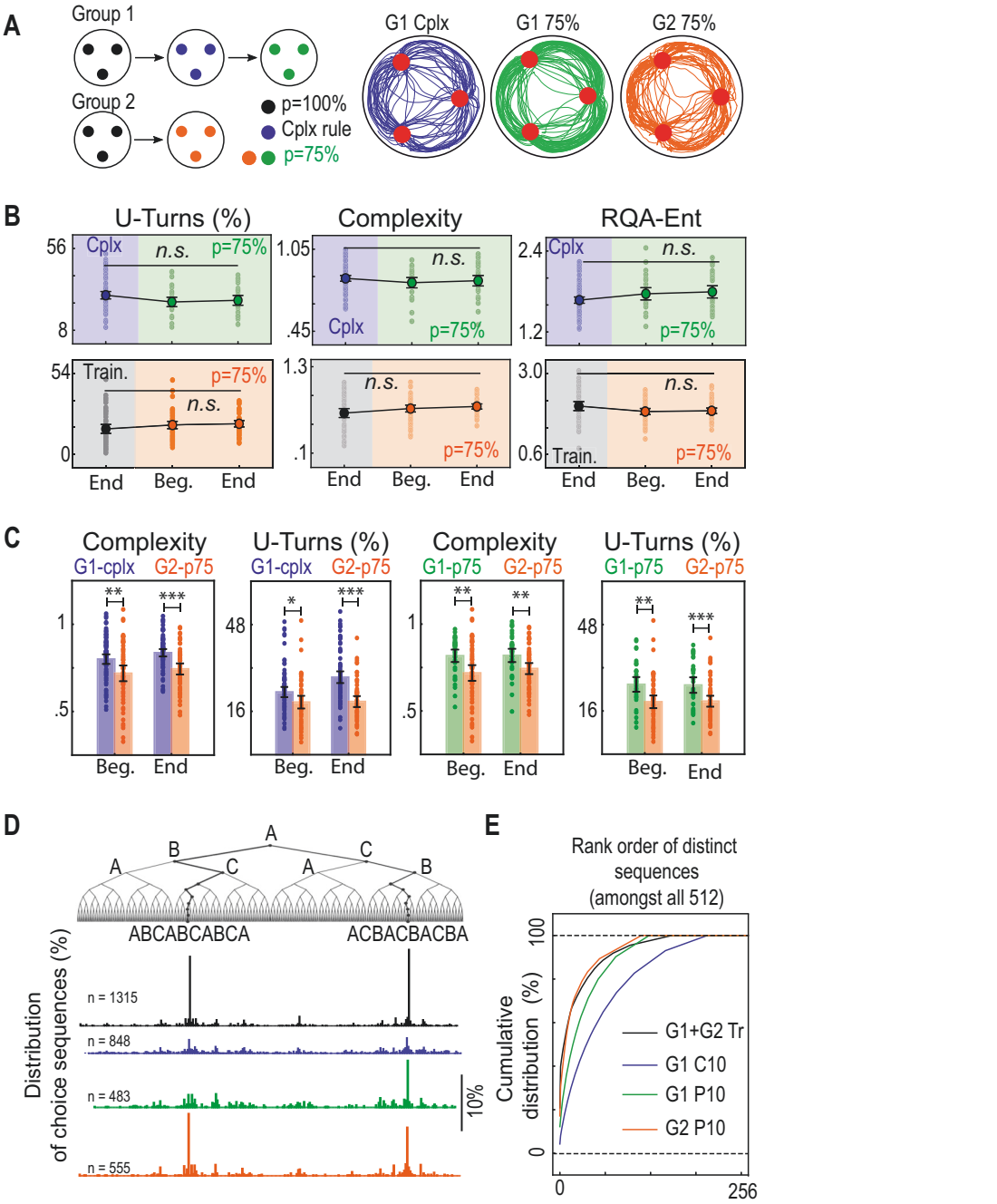


Fig. 4 Comparison of mice behavior under the complexity condition and a probabilistic condition. a Experimental setup and typical trajectories under the two conditions. For a first group of mice (G1), the complexity condition was followed by a probabilistic condition. For a second group (G2), the probabilistic condition was experienced right after training. Under the probabilistic condition all targets were rewarded with a 75% probability. **b** G1 and G2 mice behavior in the probabilistic setting compared to the end of the preceding condition (resp. complexity and training). The U-turn rate, *NLZcomp* complexity, and RQA *ENT* measure remain unchanged for G1 (pooled "end vc", "beg. p75" and "end p75", U-turn rate, $H = 4.22$, $p = 0.120$, Complexity, $H = 0.90$, $p = 0.637$, RQA *ENT*, $H = 4.57$, $p = 0.101$, Kruskal–Wallis test) and for G2 (pooled "end tr", "beg. p75", and "end p75", U-turn rate, $H = 5.68$, $p = 0.058$, Complexity, $H = 4.10$, $p = 0.128$, RQA *ENT*, $H = 2.66$, $p = 0.073$, Kruskal–Wallis test). **c** Comparison of G2 behavior in the probabilistic setting with G1 behavior under the complexity and the probabilistic conditions. G1 mice exhibit higher sequence complexity and U-turn rate than G2 under both the complexity condition (G1-cplx versus G2-p75, Complexity, pooled "beg", $t(136) = 2.99$, $p = 0.003$, pooled "end", $t(136) = 4.72$, $p = 7.10^{-6}$, Welch *t*-test, U-turn, pooled "beg", $U = 2866.5$, $p = 0.015$, pooled "end", $U = 3493$, $p = 10^{-7}$, Mann–Whitney test) and the probabilistic condition (G1-p75 versus G2-p75, Complexity, pooled "beg", $U = 1375$, $p = 0.005$, Mann–Whitney test, pooled "end", $t(91) = 2.92$, $p = 0.004$, *t*-test, U-turn, pooled "beg", $U = 1478$, $p = 0.0003$, pooled "end", $U = 1424$, $p = 0.001$, Mann–Whitney test). $N = 80$ for G1-cplx, $N = 36$ for G1-p75, $N = 54$ for G2-p75. **d** Distribution of subsequences of length 10 performed by G1 and G2 animals in the last sessions of the training, the complexity (same as in Fig. 3b), and the probabilistic conditions. **e** Cumulative distribution of ranked patterns of length 10.

aware of the existence of these possible strategies. Mice had no way of discovering that a 100% success rate could be obtained with an optimal sequence learning before ever reaching such a level of performance. In fact, we postulated that the optimal behavior would be too difficult to implement by the animals and that they shall turn to random selection instead. Overall, our results indicate that this is the case, as we found no evidence of sequence memorization nor any behavioral pattern that might have been used by mice as a heuristic to solve the complex task.

Whether and how the brain can generate random patterns has always been puzzling[23]. In this study, we addressed two fundamental aspects in this matter: the implication of memory processes and the dependence upon external (environmental) factors. Regarding memory, one hypothesis holds that in human, the process of generating random patterns leverages memory[24], to ensure the equality of response usage for example[25]. Such a procedure could indeed render choices uniformly distributed but is also very likely to produce structured sequences (i.e. dependence upon previous choices). A second hypothesis suggests that the lack of memory may help eliminate counterproductive biases[26,27]. Our experiments revealed neither sequence learning nor structure, thus supporting the latter hypothesis and the notion that the brain is able to effectively achieve high variability by suppressing biases and structure, at least in some contexts. The second aspect is the degree of dependence upon external, environmental factors. Exploration and choice variability are generally studied by introducing stochasticity and/or volatility in environmental outcomes[16–19]. However, such conditions make it difficult to interpret the animal's strategy and to know whether the observed variability in the mouse choice is inherited or not from the statistics of the behavioral task. In this work, we took a step further toward understanding the processes underlying the generation of variability per se, independently from environmental conditions. Confronted with a deterministic task which yet favors complex choice sequences, mice avoided repetitions by engaging in a behavioral mode where decisions were random and independent from their reward history. Animals adaptively tuned their decision-making parameters to increase choice randomness, which suggests an internal process of randomness generation.

Methods

Animals. Male C57BL/6J (WT) mice obtained from Charles Rivers Laboratories France (L'Arbresle Cedex, France) were used. Mice arrived to the animal facility at 8 weeks of age, and were housed individually for at least 2 weeks before the electrode implantation. Behavioral tasks started one week after implantation to ensure full recovery. Since intracranial self-stimulation (ICSS) does not require food deprivation, all mice had ad libitum access to food and water except during behavioral sessions. The temperature (20–22 °C) and humidity was automatically controlled and a circadian light cycle of 12/12 h light–dark cycle (lights on at 8:30 a.m.) was maintained in the animal facility. All experiments were performed during the light cycle, between 09:00 a.m. and 5:00 p.m. Experiments were conducted at Sorbonne University, Paris, France, in accordance with the local regulations for animal experiments as well as the recommendations for animal experiments issued by the European Council (directives 219/1990 and 220/1990).

ICSS. Mice were introduced into a stereotaxic frame and implanted unilaterally with bipolar stimulating electrodes for ICSS in the medial forebrain bundle (MFB, anteroposterior = 1.4 mm, mediolateral = ±1.2 mm, from the bregma, and dorsoventral = 4.8 mm from the dura). After recovery from surgery (1 week), the efficacy of electrical stimulation was verified in an open field with an explicit square target (side = 1 cm) at its center. Each time a mouse was detected in the area (D = 3 cm) of the target, a 200-ms train of twenty 0.5-ms biphasic square waves pulsed at 100 Hz was generated by a stimulator. Mice self-stimulating at least 50 times in a 5 min session were kept for the behavioral sessions. In the training condition, ICSS intensity was adjusted so that mice self-stimulated between 50 and 150 times per session at the end of the training (ninth and tenth session), then the current intensity was kept the same throughout the different settings.

Training session. Experiment were performed in a 1-m diameter circular open-field with three explicit location on the floor. Experiments were performed using a video camera, connected to a video-track system, out of sight of the experimenter. A home-made software (Labview National instrument) tracked the animal, recorded its trajectory (20 frames per s) for 5 min and sent TTL pulses to the ICSS stimulator when appropriate (see below). Mice were trained to perform a sequence of binary choices between the two out of three target locations (A, B, and C) associated with ICSS rewards. In the training phase all target had a 100% probability of reward.

Complexity task. In the complexity condition, reward delivery was determined by an algorithm that estimated the grammatical complexity of animals' choice sequences. More specifically, at a trial in which the animal was at the target location A and had to choose between B and C, we compared the LZ-complexity[21] of the subsequences comprised of the nine past choices and B or C (last nine choices concatenated with the two options). Both choices were rewarded if those subsequences were of equal complexity. Otherwise, only the option making the subsequence of highest complexity was rewarded. Giving that the reward delivery is deterministic, the task can be seen as a decision tree in which some paths ensure 100% rewards. From a local perspective, for each trial, the animal has either 100 or 50% chance of reward; resp. if the evaluated subsequences of size 10 have equal or unequal complexities. Considering all these possible sequences, 75% of the trials would be rewarded if animals were to choose randomly.

Measures of choice variability. Two measures of complexity were used to analyze mouse behavior. First, the normalized LZ-complexity (referred to as NLZcomp or simply complexity throughout the paper) which corresponds to the LZ-complexity divided by the average LZ-complexity of 1000 sequences of the same length generated randomly (a surrogate) with the constraint that two consecutive characters could not be equal, as in the experimental setup. NLZcomp is small for highly repetitive sequence and is close to 1 for uncorrelated, random signals. Second, the entropy of the frequency distribution of the diagonal length (noted RQA *ENT*), taken from recurrence quantification analysis (RQA). RQA is a series of methods in which the dynamics of complex systems are studied using recurrence plots (RP)[28,29] where diagonal lines illustrate recurrent patterns. Thus, the entropy of diagonal lines reflects the deterministic structure of the system and is smaller for uncorrelated, random signals. RQA was measured using the Recurrence-Plot Python module of the "pyunicorn.timeseries" package.

Computational models. The task was represented as a Markov Decision Process (MDP) with three states $s \in \{A, B, C\}$ and three actions $a \in \{GoToA, GoToB, GoToC\}$, respectively, corresponding to the rewarded locations and the transitions between them. State-action values $Q(s, a)$ were learned using the Rescorla-Wagner rule[22]:

$$\Delta Q(\mathbf{s_t}, a_t) = \alpha(\mathcal{U}_{t+1} - Q(\mathbf{s_t}, a_t)) \quad (1)$$

where $\mathbf{s_t} = [S_t, S_{t-1}, \ldots, S_{t-m}]$ is the current state, which may include the memory of up to the mth past location, a_t the current action, α the learning rate and U the utility function defined as follows:

$$\mathcal{U}_{t+1} = \begin{cases} (1-\kappa).r_{t+1} & \text{if } s_{t+1} = s_{t-1} \\ r_{t+1} & \text{otherwise} \end{cases} \quad (2)$$

where r is the reward function and κ the U-turn cost parameter modeling the motor cost or any bias against the action leading the animal back to its previous location. The U-turn cost was necessary to reproduce mouse stereotypical trajectories at the end of the training phase (see Supplementary Fig. 2).

Action selection was performed using a softmax policy, meaning that in state $\mathbf{s_t}$ the action a_t is selected with probability:

$$P(a_t|\mathbf{s_t}) = \frac{e^{Q(\mathbf{s_t}, a_t)/\tau}}{\sum_a e^{Q(\mathbf{s_t}, a_t)/\tau}} \quad (3)$$

where τ is the temperature parameter. This parameter reduces the sensitivity to the difference in actions values thus increasing the amount of noise or randomness in decision-making. The U-turn cost κ has the opposite effect since it represents a behavioral bias and constrains choice randomness. We refer to the hyperparameter defined as $\rho = \tau/\kappa$ as the randomness parameter.

In the version referred to as *BasicRL* (see Supplementary Fig. 2), we did not include any memory of previous locations nor any U-turn cost. In other words, $m = 0$ (i.e. $\mathbf{s_t} = [s_t]$) and $\kappa = 0$.

To manipulate state representation ambiguity (see Fig. 2), each of the locations {A, B, C} could be represented by $n \geq 1$ states. For simplicity, we used $n = 1$, 2, and 3 for all locations for what we referred to as 'null', 'low, and 'med' levels of ambiguity. This allowed us to present a proof of concept regarding the potential impact of using a perfect state representation in our model.

Model fitting. The main model-fitting results presented in this paper were obtained by fitting the behavior of the mice under training and complexity conditions session by session independently. This process aimed to determine which values of the two hyperparameters m and $\rho = \tau/\kappa$ make the model behave as mice in terms of success rate (i.e. percentage of rewarded actions) and complexity (i.e.

variability of decisions). Our main goal was to decide between the two listed strategies that can solve the task: repeating rewarded sequences or choosing randomly. Therefore, we momentarily put aside the question of learning speed and only considered the model behavior after convergence. α was set to 0.1 in these simulations.

Hyperparameters were selected through random search[30] (see ranges listed in Supplementary Table 1). The model was run for 2.10^6 iterations for each parameter set. The fitness score with respect to mice average data at each session was calculated as follows:

$$\text{fitness} = 1 - D_{\text{session}} \tag{4}$$

$$\text{with } D_{\text{session}} = \frac{1}{2}\left(\left|\hat{S} - \bar{S}\right| + \left|\hat{C} - \bar{C}\right|\right) \tag{5}$$

where $\bar{S}$ and $\bar{C}$ are the average success rate and complexity in mice respectively and $\hat{S}$ and $\hat{C}$ the model success rate and complexity – all the four $\in [0, 1]$. Simulations were long enough for the learning to converge. Thus, instead of multiple runs for each parameter set, which would have been computationally costly, $\hat{S}$ and $\hat{C}$ were averaged over the last 10 simulated sessions. We considered that 1 simulated session = 200 iterations, which is an upper bound for the number of trials performed by mice in one actual session.

Since mice were systematically rewarded during training, their success rate under this condition was not meaningful. Thus, to assess the ability of the model to reproduce stereotypically circular trajectories in the last training session, we replaced $\hat{S}$ and $\bar{S}$ in Eq. (5) by $\hat{U}$ and $\bar{U}$ representing the average U-turn rates for mice and for the model respectively.

Additional simulations were conducted with two goals: (1) test whether one single parameter set could fit mice behavior without the need to change parameter values over sessions, (2) test the influence of state representation ambiguity on memory use in the computational model. Therefore, each simulation attempted to reproduce mice behavior from training to the complexity condition. Hence, the learning rate α was optimized in addition to the previously mentioned m and $\rho = \tau/\kappa$ hyperparameters (see ranges listed in Supplementary Table 1). Each parameter set was tested over 20 different runs. Each run is a simulation of 4000 iterations, which amounts to 10 training sessions and 10 complexity sessions since simulated sessions consist of 200 iterations. The fitness score was computed as the average score over the last training session and the 10 complexity sessions using Eqs. (4) and (5). Using a grid search ensured comparable values for different levels of ambiguity ('null', 'low, and 'med'; see previous section). Given the additional computational cost induced by higher ambiguity levels, we gradually decreased the upper bound of the memory size range in order to avoid long and useless computations in uninteresting regions of the search space.

A sample code for the model fitting procedure is publicly available at https://zenodo.org/record/2564854#.Xe07NB-YUpg (see ref. [31]).

Markov chain analysis. Markov chain analysis allows to mathematically describe the dynamic behavior of the system, i.e. transitions from one state to another, in probabilistic terms. A process is a first-order Markov chain (or more simply Markovian) if the transition probability from state A to a state B depends only on the current state A and not on the previous ones. Put differently, the current state contains all the information that could influence the realization of the next state. A classical way to demonstrate that a process is Markovian is to show that the sequence cannot be described by a zeroth-order process, i.e. that P(B|A) ≠ P(B), and that the second-order probability is not required to describe the state transitions, i.e. that P(B|A) = P(B|AC).

In this paper, we analyzed the 0th, 1st, and 2nd order probabilities in sequences performed by each mouse in the last session of the complex condition (c10). Using the targets A, B, and C as the Markov chain states would have provided a limited amount of data. Instead, we described states as movements to the left (L) or to the right (R) thereby obtaining larger pools of data (e.g. R = {A → B, B → C, C → A}) and a more compact description (e.g. two 0th order groups instead of three). The probability of a transition (i.e. to the left or to the right, Fig. 3a) is different from the probability of the same transition given the previous one (p(L) versus P(L|L), $t(25) = -7.86$, $p = 3.10–8$, p(L) versus P(L|R), $t(25) = 7.57$, $p = 6.10–8$, $p(R)$ versus P(R|R), $t(25) = -7.57$, $p = 6.10–8$, p(R) versus P(R|L), $t(25) = 7.86$, $p = 3.10–8$, paired t-test). However, the probability given two previous transitions is not different from the latter (p(L|L) versus P(L|LL), $t(25) = 1.36$, $p = 0.183$, p(L|L) versus P(L|LR), $t(25) = -1.66$, $p = 0.108$, p(L|R) versus P(L|RL), $t(25) = -0.05$, $p = 0.960$, p(L|R) versus P(L|RR), $t(25) = -0.17$, $p = 0.860$, p(R|R) versus P(R|RR), $t(25) = 0.17$, $p = 0.860$, p(R|R) versus P(R|RL), $t(25) = 0.05$, $p = 0.960$, p(R|L) versus P(R|LR), $t(25) = 1.66$, $p = 0.108$, p(R|L) versus P(R|LL), $t(25) = -1.36$, $p = 0.183$, paired t-test).

To assess the influence of rewards on mouse decisions when switching to the complexity condition (i.e. win-stay-lose-switch strategy), we also compared the probability of going forward P(F) or backward P(U) with the conditional probabilities given the presence or absence of reward (e.g. P(F|rw) or P(U|rw)). In this case, F = R → R, L → L and U = R → L, R → L. These probabilities (Fig. 3c) were not different from the conditional probabilities given that the previous choice was rewarded or not (c01, P(F), P(F|rw) and P(F|unrw), $H = 2.93$, $p = 0.230$, P(U), P(U|rw) and P(U|unrw), $H = 1.09$, $p = 0.579$, c02, P(F), P(F|rw) and P(F|unrw), H = 1.08, p = 0.581, P(U), P(U|rw) and P(U|unrw), $H = 0.82$, $p = 0.661$, c10, P(F), P(F|rw) and P(F|unrw), $H = 0.50$, $p = 0.778$, P(U), P(U|rw) and P(U|unrw), $H = 0.50$, $p = 0.778$, Kruskal–Wallis test). In the latter analysis, we discarded the data in which ICSS stimulation could not be associated with mouse choices with certainty, due to time lags between trajectory data files and ICSS stimulation data files, or due to the animal moving exceptionally fast between two target locations (<1 s)

Analysis of subsequences distribution. All patterns starting by A were extracted and pooled from the choice sequences of mice in the last sessions of the three conditions (training, complexity, probabilistic). The histograms represent the distribution of these patterns following the decision tree structure. In other words, two neighbor branches shared the same prefix.

Bayesian model comparison. Bayesian model comparison aims to quantify the support for a model over another based on their respective likelihoods P($D|M$), i.e. the probability that data D are produced under the assumption of model M. In our case, it is useful to compare the fitness of the M_{ind} model fitted session by session independently from that of the model M_{con} fitted to all sessions in a continuous way. Since these models do not produce explicit likelihood measures, we used approximate Bayesian computation: considering the 15 best fits (i.e. the 15 parameter sets that granted the highest fitness score), we estimated the models' likelihood as the fraction of $(\hat{S}, \hat{C})$ pairs that were within the confidence intervals of mouse data. Then, the Bayes factor was calculated as the ratio between the two competing likelihoods:

$$B = \frac{P(D/M_{\text{ind}})}{P(D/M_{\text{con}})} \tag{6}$$

B > 3 was considered to be a substantial evidence in favor of M_{ind} over M_{con}[32].

Statistics and reproducibility. No statistical methods were used to predetermine sample sizes. Our sample sizes are comparable to many studies using similar techniques and animal models. The total number of observations (N) in each group as well as details about the statistical tests were reported in figure captions. Error bars indicate 95% confidence intervals. Parametric statistical tests were used when data followed a normal distribution (Shapiro test with $p > 0.05$) and non-parametric tests when they did not. As parametric tests, we used t-test when comparing two groups or ANOVA when more. Homogeneity of variances was checked preliminarily (Bartlett's test with $p > 0.05$) and the unpaired t-tests were Welch-corrected if needed. As non-parametric tests, we used Mann–Whitney test when comparing two independent groups, Wilcoxon test when comparing two paired groups and Kruskal–Wallis test when comparing more than two groups. All statistical tests were applied using the scipy.stats Python module. They were all two-sided except Mann–Whitney. $p > 0.05$ was considered to be statistically non-significant.

In all Figures: error bars represent 95% confidence intervals. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. n.s., not significant at $p > 0.05$[33].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study[33] are available at https://zenodo.org/record/3576423#.Xfdez-tCe3A and from the corresponding author upon reasonable request.

Code availability

A sample code for the model fitting procedure[31] is publicly available at https://zenodo.org/record/2564854#.Xe07NB-YUpg.

Received: 1 December 2019; Accepted: 20 December 2019;
Published online: 21 January 2020

References

1. Wu, H. G., Miyamoto, Y. R., Gonzalez Castro, L. N., Ölveczky, B. P. & Smith, M. A. Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nat. Neurosci.* **17**, 312–321 (2014).
2. Aronov, D., Andalman, A. S. & Fee, M. S. A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* **320**, 630–634 (2008).
3. Driver, P. M. & Humphries, D. A. *Protean behaviour*. (Oxford University Press, USA, 1988).
4. Rapoport, A. & Budescu, D. V. Generation of random series in two-person strictly competitive games. *J. Exp. Psychol. Gen.* **121**, 352–363 (1992).
5. Sutton, R. S. & Barto, A. G. *Reinforcement Learning*. (MIT Press, 1998).

6. Schultz, W. Getting formal with dopamine and reward. *Neuron* **36**, 241–263 (2002).
7. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **362**, 933–942 (2007).
8. Rao, R. P. N. Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Front. Comput. Neurosci.* **4**, 146 (2010).
9. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
10. Mansouri, F. A., Koechlin, E., Rosa, M. G. P. & Buckley, M. J. Managing competing goals - a key role for the frontopolar cortex. *Nat. Rev. Neurosci.* **18**, 645–657 (2017).
11. Grunow, A. & Neuringer, A. Learning to vary and varying to learn. *Psychonomic Bull. Rev.* **9**, 250–258 (2002).
12. Kane, G. A. et al. Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cogn. Affect Behav. Neurosci.* **17**, 1–11 (2017).
13. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
14. Karlsson, M. P., Tervo, D. G. R. & Karpova, A. Y. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science* **338**, 135–139 (2012).
15. Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S. & Wyart, V. Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nat. Neurosci.* **441**, 876–12 (2019).
16. Naudé, J. et al. Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. *Nat. Neurosci.* **19**, 471–478 (2016).
17. Cinotti, F. et al. Dopamine regulates the exploration-exploitation trade-off in rats. 1–36, https://doi.org/10.1101/482802 (2019).
18. Lee, D., Conroy, M. L., McGreevy, B. P. & Barraclough, D. J. Reinforcement learning and decision making in monkeys during a competitive game. *Cogn. brain Res.* **22**, 45–58 (2004).
19. Tervo, D. G. R. et al. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* **159**, 21–32 (2014).
20. Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7**, 404–410 (2004).
21. Lempel, A. & Ziv, J. On the complexity of finite sequences. *IEEE Trans. Inf. Theory* **22**, 75–81 (1976).
22. Rescorla, R. A. & Wagner, A. R. *A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement.* In (eds AH. Black & W.F. Prokasy), Classical conditioning II: current research and theory. 64–99 (Appleton-Century-Crofts, New York, 1972).
23. Glimcher, P. W. Indeterminacy in brain and behavior. *Annu Rev. Psychol.* **56**, 25–56 (2005).
24. Towse, J. N. & Cheshire, A. Random number generation and working memory. *Eur. J. Cogn. Psychol.* **19**, 374–394 (2007).
25. Oomens, W., Maes, J. H. R., Hasselman, F. & Egger, J. I. M. A time series approach to random number generation: using recurrence quantification analysis to capture executive behavior. *Front. Hum. Neurosci.* **9**, 319 (2015).
26. Wagenaar, W. Generation of random sequences by human subjects: a critical survey of literature. *Psychological Bull.* **77**, 65–72 (1972).
27. Maes, J. H. R., Eling, P. A. T. M., Reelick, M. F. & Kessels, R. P. C. Assessing executive functioning: on the validity, reliability, and sensitivity of a click/point random number generation task in healthy adults and patients with cognitive decline. *J. Clin. Exp. Neuropsychol.* **33**, 366–378 (2011).
28. Marwan, N., Romano, M. C., Thiel, M. & Kurths, J. Recurrence plots for the analysis of complex systems. *Phys. Rep.* **438**, 237–329 (2007).
29. Faure, P. & Lesne, A. Recurrence plots for symbolic sequences. *Int. J. Bifur. Chaos* **20**, 1731–1749 (2010).
30. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
31. Belkaid, M. Code for basic q-learning model fitting, https://doi.org/10.5281/zenodo.2564854 (2019).
32. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
33. Belkaid, M. et al. Mice adaptively generate choice variability in a deterministic task - behavioral data. https://doi.org/10.5281/zenodo.3576423 (2019).

Acknowledgements

This work was supported by the Centre National de la Recherche Scientifique CNRS UMR 8246 et UMR 7222, the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02, the Foundation for Medical Research (FRM, Equipe FRM EQU201903007961 to P.F), ANR (ANR-16 Nicostress to PF), the French National Cancer Institute Grant TABAC-16-022 (to P.F.). P.F. team is part of the École des Neurosciences de Paris Ile-de-France RTRA network and member of LabEx Bio-Psy.

Author contributions

P.F. designed the behavioral experiment. P.F., E.B., R.D.C., M.D., E.D., T.A.Y., B.H. and M.C. performed the behavioral experiments, P.F. and M.B. analyzed the behavioral data. M.B. developed the computational model, S.D. developed some acquisition tools. J.N. and O.S. contributed to modeling studies and data analysis. M.B., P.F., J.N. and O.S. wrote the paper with inputs from A.M.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42003-020-0759-x.

Correspondence and requests for materials should be addressed to P.F.

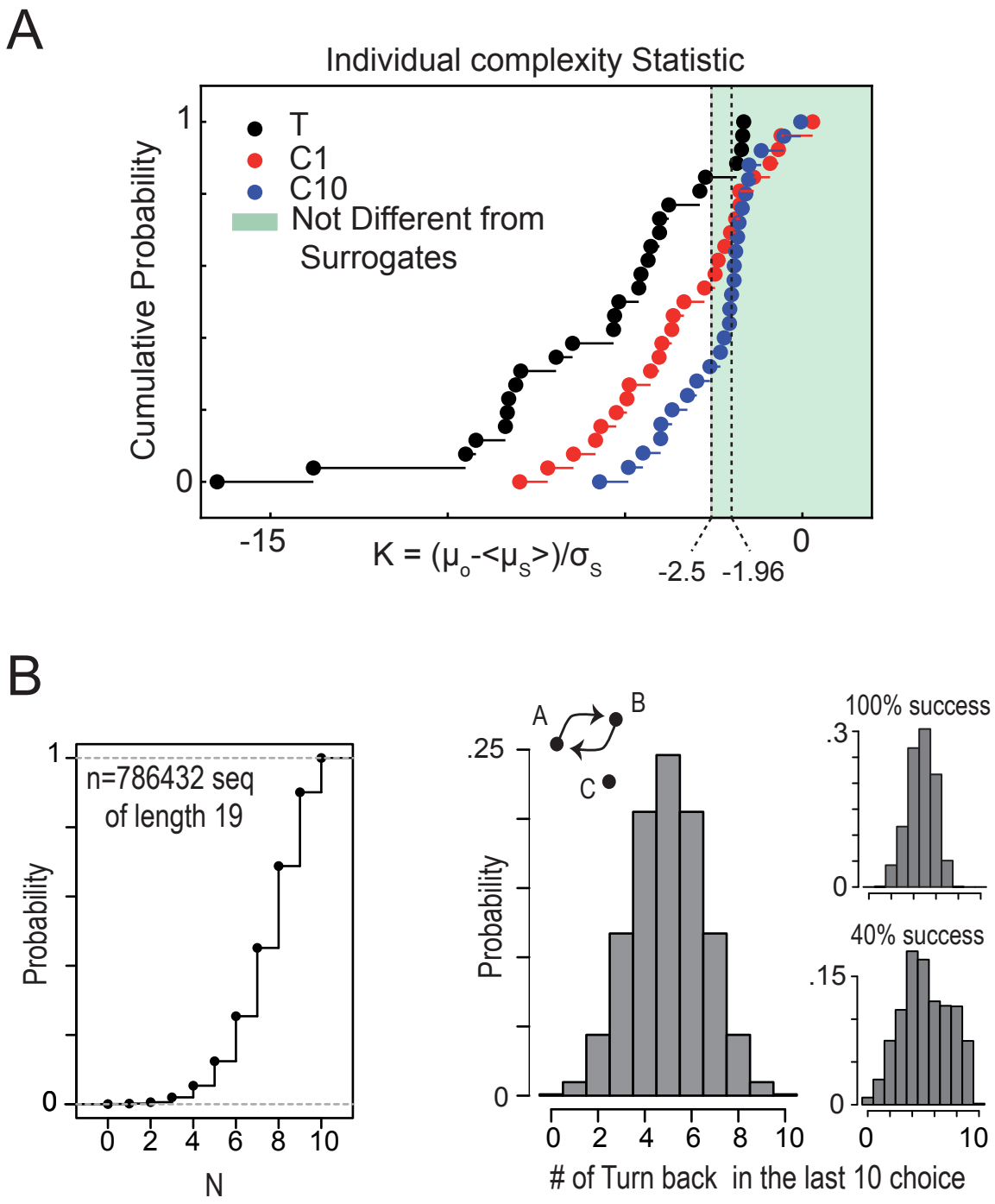
Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

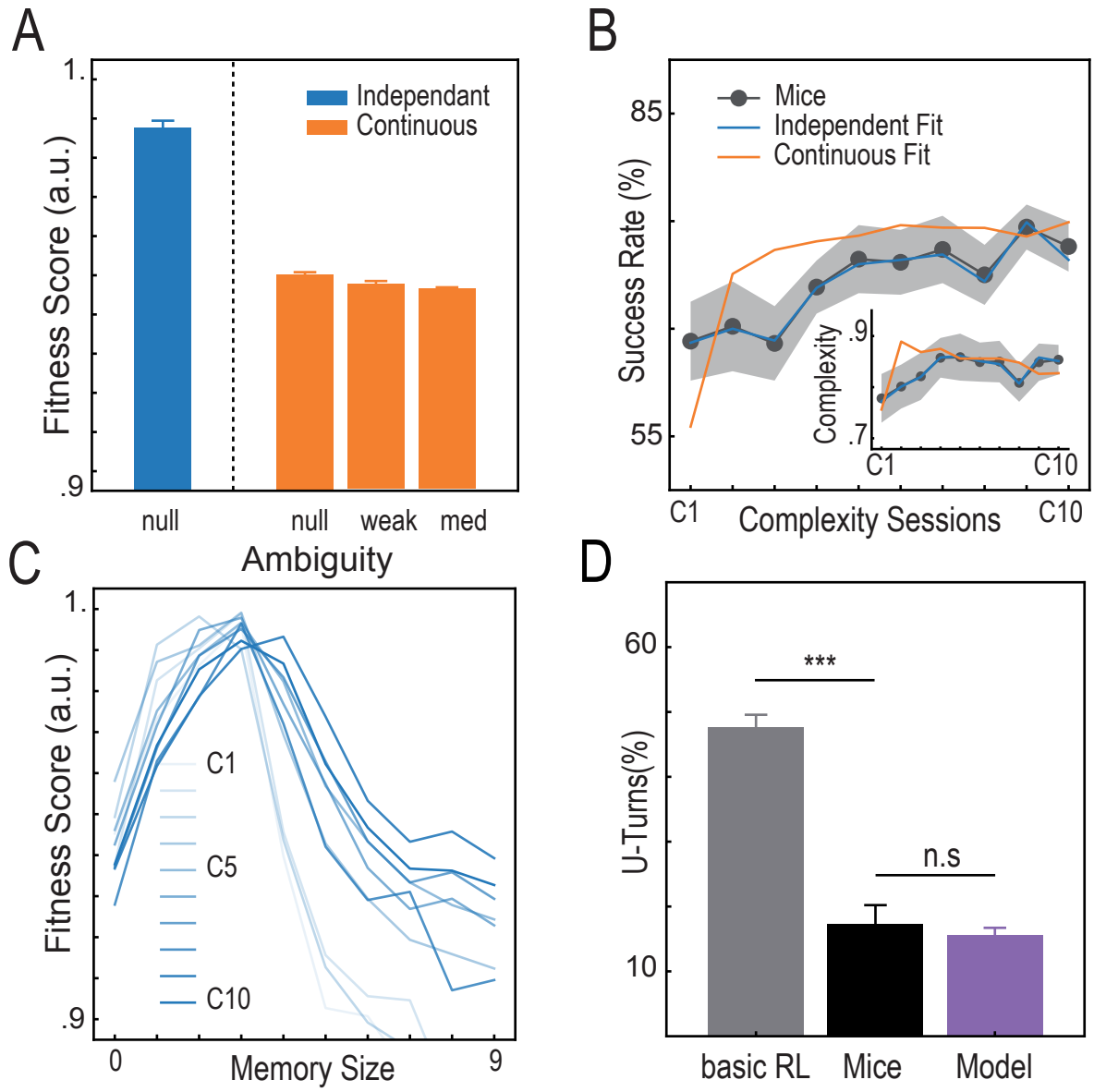
© The Author(s) 2020

Supplementary information : Mice adaptively generate choice variability in a deterministic task , Marwen Belkaid, Elise Bousseyrol, Romain Durand-de Cuttoli, Malou Dongelmans, Etienne K. Duranté , Tarek Ahmed Yahia, Steve Didienne, Bernadette Hanesse, Maxime Come, Alexandre Mourot, Jérémie Naudé, Olivier Sigaud et Philippe Faure *



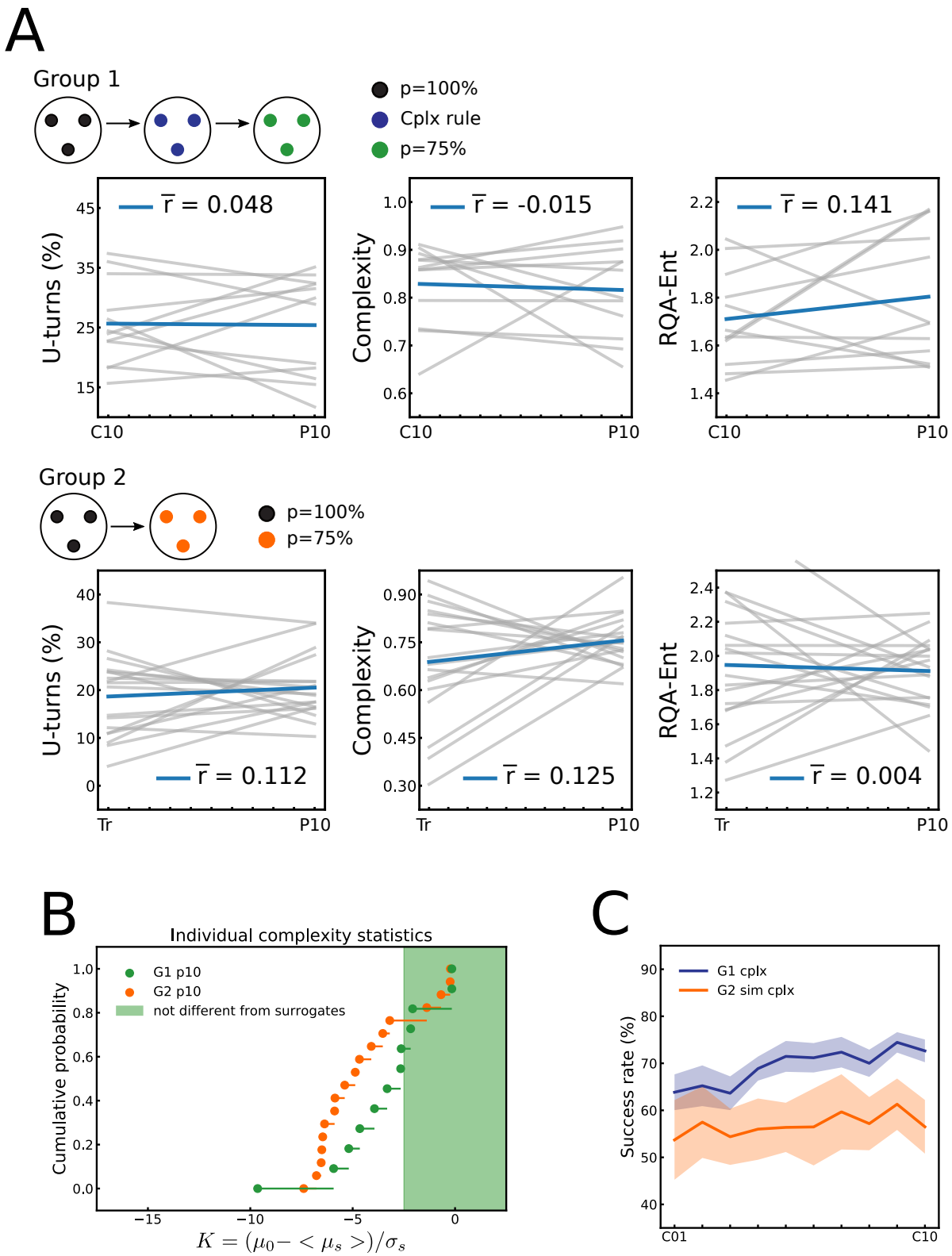
Supplementary Figure 1

Supplementary Figure 1: A) Individual comparison with random sequence of choices: Cumulative distribution of the K parameters calculated for each sequence of choices in sessions Training (T), C1 and C10. For each experimental sequence of choices, the term $K=(\mu_o-\langle\mu_s\rangle)/\sigma_s$ was calculated, where μ_o is the complexity of the original data, $\langle\mu_s\rangle$) and σ_s are the mean and standard deviation of the complexity of the surrogate series (i.e. for each experimental series, 1000 random sequences of the same length in which two consecutive elements could not be equal). We then tested the hypothesis that each original set was different from surrogates. Assuming Gaussian statistics, a limit of K=-2.5 and -1.96 indicates respectively a confidence of 99.4% and 95% that $\mu_o \geq \mu_s$. We found that 65.4% of mouse sequences in session C10 were not different from surrogates with a confidence of 99.4%. N = 27 in all sessions except C10 where N = 26. **B) Theoretical number of rewards and U-turns in the last 10 choices of a sequence of length 19:** *Left*: Cumulative distribution of the number of rewards obtained in the last 10 choices of the total number of sequences of length 19. Middle: Histogram of the number of U-turns in the last 10 choices. *Right*: Histogram of the number of U-turns in all possible sequences with 100% of reward (top) or 40% of reward (bottom). The optimal U-turn rate to maximize rewards is 50%.



Supplementary Figure 2

Supplementary Figure 2: Model fitting results for different variants of the model. A), **B)** and **C)** Comparison of fitness scores with independent and continuous model fitting procedures. 'Independent' refers to the session-by-session hyperparameter optimization, 'continuous' to the optimization of the model for all sessions in a row (see 'Methods'). **A)** Comparison of the fitness score obtained by the N = 15 best fits in the independent and continuous fitting procedure. In continuous fitting, three levels of state representation ambiguity are shown (see 'Methods'). Error bars represent 95% confidence intervals. **B)** Success rates and complexity levels obtained in simulations in comparison with those obtained by mice. The shaded area represents the 95% confidence interval for mice. Only best fits are represented for model simulations (average of 20 runs). The model fitted for each session independently follows the evolution of behavioral data by adapting the exploration hyperparameter. By contrast, fitting all sessions with one single parameter set fails to reproduce the same evolution. In particular, the model exhibits a radical increase in success rate and complexity between sessions C01 and C02 as opposed to a more gradual progress in mice. **C)** Best fitness scores obtained with each of the tested memory sizes session by session. Smaller memories fit better. **D)** Comparison between the U-turn rates achieved by a basic RL algorithm, by our model and by mice in the training phase. A basic RL algorithm is unable to reproduce the stereotypical circular trajectories observed in mice during and after training; see also Figure 1B for mouse trajectories). Indeed, with equal probabilities of reward at all targets, this algorithm learns equal state-action values and randomly chooses between them. Discounting the reward function by a U-turn cost representing previous locations in the state vector (see 'Methods') makes the model capable to generate the same percentage of U-turns as mice under the training condition.



Supplementary Figure 3

Supplementary Figure 3: Absence of change in terms of U-turn rate, complexity and RQA entropy in the probabilistic condition compared to the previous condition. *Top*: Animals from Group 1. *Bottom*: Animals from Group 2. Grey lines represent linear regressions of the three measures for individual mice and the blue line represents the average. **B) Individual comparison with random sequence of choices:** Cumulative distribution of the K parameters calculated for each sequence of choices in the last sessions under the probabilistic rule p10 for G1 and G2 animals. For each experimental sequence of choices, the term $K=(\mu_o-<\mu_s>)/\sigma_s$ was calculated, where μ_o is the complexity of the original data, $<\mu_s>$) and σ_s are the mean and standard deviation of the complexity of the surrogate series (i.e. for each experimental series, 1000 random sequences of the same length in which two consecutive elements could not be equal). We then tested the hypothesis that each original set was different from surrogates. Assuming Gaussian statistics, a limit of K=-2.5 and -1.96 indicates respectively a confidence of 99.4% and 95% that $\mu_o \geq \mu_s$. We found that 33.3% of G1 mouse sequences and 22.3% of G2 mouse sequences were not different from surrogates with a confidence of 99.4%. N = 12 for G1, N = 18 for G2. **B) Success rate of the sequences performed by G2 animals if they were performed under the complexity rule:** We simulated the reward delivery rule of the complexity condition against the sequences performed by G2 mice under the probabilistic condition. Under these conditions, the animals would have reached a significantly lower success rate starting from the third session.

Supplementary Table 1: Hyperparameter ranges used in random search in independent, session-by-session model fitting ($N_{samples} = 6000$) and in grid search in continuous, all-sessions model fitting.

Hyperparameter random search ranges

Label	Range	Step	Description
m	$[0, 9]$	1	Memory size
τ	$[1, 20]$	*continuous*	Softmax temperature
κ	$[0, 1]$	*continuous*	U-turn cost

Hyperparameter grid search ranges

Label	Range	Step	Description
m	$[0, 9]$	1	Memory size (no ambiguity)
	$[0, 7]$	1	Memory size (low ambiguity)
	$[0, 5]$	1	Memory size (medium ambiguity)
α	$2^{-i}, i \in [0, 10]$	1	Learning rate
τ	$2^{i}, i \in [-4, 4]$	1	Softmax temperature
κ	$[0.5, 0.95]$	0.05	U-turn cost

Chapitre 7 : Discussion générale

Au cours de ce travail de thèse, j'ai questionné les rôles des trois structures du circuit mésocortical dans la prise de décision chez la souris : la VTA, l'OFC et le PFC. Toutes trois ont été reliées dans la littérature aux comportements orientés vers un but, aux récompenses et à l'exploration. En particulier, je me suis intéressée aux différents encodages que réalisent ces structures dans un environnement probabiliste, et à leurs interactions dynamiques au cours d'une séquence de choix. Ces choix sont dits auto-générés par l'animal, c'est-à-dire qu'ils ne sont pas déclenchés par un indice explicite de l'environnement. J'ai également questionné le degré de dépendance des décisions des animaux aux facteurs externes spécifiés par l'environnement et en particulier la statistique associée à l'obtention des récompenses. L'objectif général a été de comprendre l'émergence intrinsèque d'exploration et de variabilité dans les choix.

Ces études ont été réalisées dans une tâche dans laquelle les animaux effectuent des choix de manière spontanée. Les animaux se déplacent pour obtenir des ICSS dont l'obtention constitue le résultat de leurs actions. L'utilisation d'ICSS pour le conditionnement des souris présente l'avantage de ne pas être soumis aux variations motivationnelles. En permettant directement la libération de DA dans les structures cibles du circuit méso-cortico-limbique, ce type de récompense s'affranchit des mécanismes internes de satiété ou d'ennui. A intensité constante et après apprentissage, les souris effectuent un nombre suffisamment grand de choix pour étudier de manière robuste : i) les variations fines de leur comportement, ii) leurs choix en contexte probabiliste ou complexe, et iii) les dynamiques d'activités mésocorticales. La quantification de leurs choix a été complétée par l'utilisation de modèles computationnels permettant d'interpréter les stratégies de choix utilisées par les animaux, et en dégageant notamment des paramètres les influençant.

I) Activités prédictives ou causales

Selon la vision neuroéconomique de la prise de décision, plusieurs traitements d'informations ou computationnels sont réalisés au niveau neuronal sans lesquels un choix ne saurait aboutir. Ces traitements concernent à la fois la représentation des paramètres du problème décisionnel (récompenses, incertitude, délai, facteurs internes, comparaison de valeurs subjectives, etc.), la sélection d'une stratégie, et la réalisation de l'action (initiation du plan moteur, énergie du mouvement, etc.) (Rangel et al., 2008). L'étude des bases neurales des processus de décision requiert donc de confronter les signaux électrophysiologiques recueillis à ces différentes computations possibles pour les interpréter au mieux. En particulier, il est important de se questionner sur la nature causale des motifs (*patterns*) d'activité que l'on analyse, s'ils reflètent de simples prédictions passives de stimuli environnementaux, ou s'ils sont causalement impliqués dans la politique de choix des animaux, ou leur exécution.

Dans la première étude de cette thèse (Article 1), nous avons observé une augmentation transitoire dans l'activité des neurones DAergiques de la VTA au moment où les animaux se dirigent vers un point associé à une récompense. Cette activité repose sur l'émission synchrone de bouffées de PA. Nous avons également observé une augmentation de puissance des oscillations θ (7-14 Hz) au moment du choix, qui est précédé par une augmentation de l'activité de population dans l'OFC. Ces patterns d'activité observés après apprentissage (à la fin du cadre déterministe), peuvent constituer des prédictions dans la tâche : la décision a été prise, la récompense est donc attendue et prédite. Ce mécanisme s'opère de manière interne, sans besoin d'un stimulus prédicteur externe. Aucune activité spécifique n'est observée en revanche dans le PFC. Par ailleurs, dans le cadre probabiliste, l'activité phasique DAergique corrèle en amplitude avec la probabilité de récompense associée à la cible vers laquelle les souris se dirigent, en accord avec la théorie de la RPE quantitative (Montague et al., 2004; Schultz, 1998; Sutton and Barto, 1998).

Nous avons également observé des signaux d'erreurs dans les trois structures d'intérêt, lors d'omission ou d'obtention inattendue de récompense. En effet, lors de l'omission inattendue de l'ICSS au début du cadre probabiliste, les neurones DAergiques de la VTA présentent une diminution d'activité, alors que celle-ci augmente si l'ICSS est délivrée de manière inattendue avant apprentissage. Ainsi, les neurones DAergiques signalent bien des erreurs de prédictions de récompenses en lien avec les contingences connues de

l'environnement. Par ailleurs, l'omission inattendue de l'ICSS induit une augmentation de puissance dans la gamme δ (3-6 Hz) dans le PFC, tandis que le déclenchement inattendu d'ICSS provoque une augmentation de l'activité de population dans l'OFC et le PFC, ainsi qu'une désynchronisation des oscillations θ de l'EFP dans l'OFC. Le PFC n'étant pas particulièrement actif après apprentissage en cadre déterministe, ces résultats suggèrent que son implication pourrait être spécifique à des résultats incertains.

Une première question se pose et sera l'objet d'une première partie de la discussion. En effet, que certaines activités dans la VTA, l'OFC et le PFC correspondent à des signaux attendus d'apprentissage et de prédiction ne signifie pas qu'elles soient utilisées comme telles. Ces motifs électrophysiologiques sont-ils des corrélats de valeurs utiles à l'apprentissage, ou bien sont-ils causalement impliqués dans la modulation du comportement en cours ?

1) Activité dopaminergique et erreur de prédiction de récompense

L'activité phasique des neurones DAergiques de la VTA a été de nombreuses fois reliées au calcul de RPE dans la littérature et est considérée comme un signal d'apprentissage par renforcement (Montague et al., 2004; Schultz et al., 1997; Steinberg and Janak, 2013; Tsai et al., 2009). Ce calcul de RPE, qui consiste en la différence de valeur entre ce qui est prédit et ce qui est obtenu (cf. Chapitre 1 et 2), représente de fait à la fois une prédiction, et une évaluation du résultat d'une action. Cela en fait un signal d'ajustement de la force d'une association action-résultat (*action-outcome*), et donc un élément clé dans l'apprentissage. Tout un pan de la littérature relie également l'OFC et le PFC à des prédictions de récompense (Lak et al., 2019), des comparaisons de valeurs (Padoa-Schioppa, 2009), ou des calculs d'états (Gershman and Uchida, 2019; Wilson et al., 2014b), et également à des signaux d'erreurs (Takahashi et al., 2009).

Le profil d'activité des neurones DAergiques observé dans cette tâche (cf. Article 1) correspond bien au formalisme de la RPE : il reflète une prédiction qui corrèle avec valeur du but, et signale des erreurs par rapport à ce qui est prédit. Il correspond donc bien à la théorie de l'apprentissage par différence temporelle (*TD learning*) (Schultz, 1998; Sutton and Barto, 1998) et au formalisme de la RPE. Cette activité phasique reflète-t-elle seulement un calcul de RPE ou peut-elle influencer le choix que l'animal s'apprête à faire ? Sur ce point, les résultats de la littérature ne sont pas tranchés. Si certains auteurs limitent le rôle de la DA à ce calcul de RPE

(Steinberg and Janak, 2013; Tsai et al., 2009), d'autres suggèrent que l'activité DA peut aussi être causalement liée à l'initiation de l'action (da Silva et al., 2018; Fischbach-Weiss et al., 2018; Wassum et al., 2012). Dans ce cas, la DA encoderait donc également une composante motivationnelle, en accord avec les théories sur le rôle de la DA dans la motivation incitatrice (Berridge, 2012).

Dans ce travail de thèse, nous avons corrélé l'activité phasique DAergique dans la tâche en début d'essai avec la durée des essais (*time to goal*) et le temps nécessaire pour atteindre la vitesse maximale dans un essai (*time to maximal speed*), deux paramètres comportementaux qui reflètent la vigueur et la motivation (cf. Article 1). L'inhibition de l'activité DAergique phasique par optogénétique provoque une diminution de la vitesse maximale dans l'essai, retarde l'initiation de l'action et augmente de fait la durée des essais. L'activité phasique DAergique est donc non seulement un signal d'apprentissage et d'ajustement du comportement (RPE), mais semble également reliée causalement à l'intensité motrice déployée pour atteindre le but, c'est-à-dire à la motivation.

Pouvons-nous supposer qu'elle pourrait aussi avoir un rôle sur le choix des animaux ? La manipulation de l'activité DAergique dans un cadre déterministe ne met pas en évidence de modification du demi-tour, suggérant que le choix (à valeur égale) n'est pas influencé. Dans ce cas, comment imaginer l'influence que la DA pourrait avoir sur les choix des animaux sous incertitude ? Nos travaux concluent que la dopamine signale la motivation déployée pour atteindre le but, mais qu'en ce qui concerne le choix, la structure de la tâche ne nous permet pas de conclure correctement, parce qu'il semble difficile de dissocier le calcul d'erreur et celui du choix en cours.

Une autre question se pose également concernant le rôle de l'activité DAergique dans le cadre complexe (cf. Article 2). Si l'activité phasique DAergique reflète un calcul de RPE et est un signal d'apprentissage, comment permet-elle l'apprentissage de choix variables en cadre complexe ? Dans ce contexte, la distribution des récompenses repose sur une règle déterministe qui consiste à ne récompenser que les séquences de choix complexes. La valeur de récompense de chacune des trois cibles est donc en moyenne identique, mais toutefois différente du cadre déterministe dans lequel les récompenses sont certaines en chaque cible. Nos résultats montrent que les souris s'adaptent à ce contexte en adoptant une stratégie de choix intégrant de l'aléatoire et ne reposant pas sur de la mémoire. Comment le signal de RPE se réarrange au moment de la transition entre le cadre déterministe et complexe ? Par ailleurs, quel serait le rôle de ce signal, qui est causalement impliqué à du renforcement par essai/erreur, lorsqu'il s'agit de générer de

l'aléatoire et de s'opposer à ce qui est renforcé par expérience ? Nous pouvons supposer que la DA ne détermine pas à elle seule ce genre de comportements, et que d'autres acteurs y participent.

2) Activité dans l'OFC et la représentation d'état

Nous avons observé dans l'OFC des oscillations θ qui se mettent en place au cours de l'apprentissage de la tâche déterministe. Néanmoins, elles ne correspondent pas à un calcul de RPE car ne sont pas modifiées lors de l'omission inattendue des récompenses. Il en va de même pour l'activité de population enregistrée dans l'OFC. Ces résultats semblent en accord avec les données de la littérature. Bien que les oscillations θ dans l'OFC aient été reliées à différentes computations liées aux récompenses (Amarante and Laubach, 2021; Miller and Wallis, 2013), l'OFC est plutôt impliqué dans les associations action-résultat flexibles (outcome) (Sharpe et al., 2019; Stalnaker et al., 2015).

Dans notre tâche, la transition au cours d'un essai, d'une augmentation d'activité de population à une augmentation de puissance dans la gamme θ a lieu au moment du départ de l'animal vers son but, et est corrélée avec le temps de résidence (dwell time). On considère que l'initiation de l'action correspond à un changement d'état de l'animal (état immobile à état dans lequel l'action est déclenchée). Ces observations sont cohérentes avec les théories impliquant l'OFC dans les représentations des états de la tâche (Sharpe et al., 2019, p. 201; Wilson et al., 2014b). Selon cette vision, l'OFC signalerait un message du type « vous êtes ici » sur la base d'informations explicites et inférées de l'environnement et des états internes de l'animal.

En revanche, la puissance θ dans l'OFC et l'activité de population ne sont pas affectées par l'introduction des premières omissions. Nous aurions effectivement pu nous attendre à observer au niveau de l'OFC des conséquences des changements de contingences. En outre, ces activités dans l'OFC ne sont pas non plus corrélées à la valeur de récompense ou à l'incertitude, ce qui contredit une partie de la littérature impliquant l'OFC dans les choix économiques (Cai and Padoa-Schioppa, 2014; Padoa-Schioppa, 2009). Ce rôle d'encodage de valeur subjective dans le choix pourrait éventuellement être plus spécifique des primates et des humains. Enfin, dans notre tâche, l'OFC ne reflète pas non plus la confiance (correspondant à l'inverse de l'incertitude) que certains auteurs relient à cette structure (Kepecs et al., 2008).

Notons néanmoins que cette interprétation quant au rôle de l'OFC ne repose que sur une approche corrélative. Pour le confirmer, une manipulation causale de cette structure pourrait être mise en place mais serait probablement compliquée à interpréter. En effet, l'inactivation de l'OFC aboutit souvent à des déficits d'apprentissage, sans pour autant être en mesure de conclure sur un rôle plus spécifique. Une autre possibilité serait de modifier les règles de la tâche (comme par exemple d'imposer un délai d'attente sur chaque point) afin de mesurer comment l'activité dans l'OFC est impactée.

3) Le cortex préfrontal et le changement de contingences

Le PFC semble, quant à lui, réagir plus spécifiquement à l'introduction d'incertitude dans la tâche, c'est-à-dire à un changement de contingences (cf. Article 1). En effet, bien qu'inactif durant l'apprentissage déterministe de la tâche, nous avons observé une augmentation de puissance δ lors des premières omissions, qui n'a pas été corrélée aux différents paramètres comportementaux (mesures de motivation ou répartition des choix) des animaux dans la tâche. Cela suggère que ce signal oscillatoire n'oriente pas causalement les actions des animaux, mais constitue plutôt un signal d'erreur, pouvant potentiellement être utilisé pour adapter les comportements aux nouvelles règles de la tâche.

Dans ce cas, nous pouvons nous questionner sur l'implication du PFC en cadre complexe. S'il détecte et signale les omissions de récompenses, et en considérant qu'une politique stochastique est un évitement actif d'une stratégie d'exploitation, est-il particulièrement impliqué dans la mise en place de séquences de choix complexes ? Ou bien au contraire, dans un contexte où il semble que les animaux utilisent une stratégie de choix stochastique, est-il plutôt désengagé ?

Un autre type de signal a été observé dans le PFC (cf. Article 1) : il s'agit d'un signal prédictif de l'incertitude identifié dans l'activité de population. Ce signal a par ailleurs été négativement relié à la durée des essais (*time to goal*), suggérant qu'il est causalement impliqué dans les choix des animaux. Puisque le PFC est considéré dans la littérature comme un orchestrateur des stratégies de choix, on peut se demander si cette activité reflète simplement un paramètre de l'environnement, ou est un signal activement décodé par le cerveau et permettant à l'animal d'orienter sa stratégie en cours ?

Les effets de l'inactivation du PFC par le muscimol seront discuté plus bas (cf. partie III) de la discussion). Notons néanmoins que cette manipulation causale ne permet pas de conclure quant aux rôles spécifiques de ces deux types d'activité (oscillations δ et activité de population). Par ailleurs, les signaux électrophysiologiques recueillis dans cette étude ne permettent pas distinguer les unités neuronales. Des enregistrements multi-unitaires dans le PFC, ainsi que dans l'OFC, permettraient d'évaluer plus spécifiquement les encodages des paramètres comportementaux dans la tâche. Cela permettrait d'orienter des expériences de manipulations causales vers certaines sous-populations neuronales, c'est-à-dire manipuler l'activité de cellules pyramidales ou interneurones GABA par optogénétique par exemple.

II) Séquence d'activités et contextes

L'utilisation de la tâche 3 points nous permet de tester comment les *patterns* d'activité neuronale identifiés contrôlent la mise en place de choix auto-générés selon le contexte de la tâche. Nous avons mis en évidence, dans l'activité en décharge et les oscillations de la VTA, de l'OFC et du PFC, différents corrélats et encodages qui sont distribués au sein d'une séquence. L'ensemble de cette séquence d'activité semble déterminer le but de chaque essai, ainsi que leur initiation et la vigueur du déplacement à mettre en œuvre.

L'OFC et le PFC sont généralement considérés comme des entrées des neurones DAergiques de la VTA (réseaux *feedforward*). Leur rôle serait de façonner l'activité de RPE des neurones DAergiques (Starkweather et al., 2018; Takahashi et al., 2009). C'est par ce biais indirect que l'OFC et le PFC exerceraient un contrôle sur l'apprentissage et les comportements. En revanche, des études de pharmacologie des récepteurs DAergiques dans les cortex préfrontaux, ou d'inhibition chemogénétique des voies mésocorticales DAergiques placent l'OFC et le PFC comme des cibles de la DA et ces voies comme directement impliquées dans la prise de décision sous incertitude (Puig and Miller, 2012; St. Onge and Floresco, 2010; Verharen et al., 2018). Ces approches ont tendance à considérer ces circuits comme un ensemble de modules effectuant des calculs indépendants et de manière séquentielle. Ces études ne permettent donc pas de conclure quant aux rôles et mécanismes d'interaction de la signalisation DAergique et des cortex préfrontaux dans la prise de décision. En outre, elles

ignorent le fait que le circuit mésocortical est davantage une boucle qu'une série *feedforward* de réseaux. C'est cet aspect que nous avons essayé de défendre dans le premier papier (cf. Article 1) et que je rediscuterai ici.

1) Des activités organisées en séquence

Dans la tâche utilisée au cours de ce travail de thèse, la séquence d'activité observée lors d'un essai se construit selon une structure reproductible d'un essai à un autre (Figure 17). Elle commence, avant que l'animal n'entame son action vers le point suivant, par une activité phasique des neurones DAergiques de la VTA accompagnée d'une augmentation de l'activité de population dans l'OFC, et suivi par une augmentation de puissance dans la gamme θ (7-14 Hz) dans l'OFC. L'activité de population de l'OFC a été corrélée à l'initiation de l'action (*dwell time*), suggérant une participation précoce de cette structure dans les décisions des animaux. La suppression de l'activité phasique DAergique par optogénétique au moment de l'initiation perturbe les oscillations θ dans l'OFC. Cela suggère que la DA jouerait un rôle dans la transformation d'une activité de population synchrone en oscillations θ, soit de manière directe, soit de manière indirecte en induisant un changement d'état de l'animal qui aboutirait ensuite à une synchronisation θ. L'activité phasique DAergique n'étant pas corrélée avec le temps de résidence sur les cibles (*dwell time*), qui constitue le temps le plus précoce de la séquence comportementale, il est possible que l'activité DAergique se place plus tardivement dans l'essai que l'activité de population de l'OFC. Cela laisserait penser à une interaction récurrente de type OFC-VTA-OFC, qui pourrait être confirmée en enregistrant l'activité dans la VTA et l'OFC simultanément sur les mêmes animaux.

En cadre probabiliste, l'activité de population dans le PFC reflétant l'incertitude attendue survient également précocement dans la séquence comportementale, avant que l'animal n'effectue son choix. En outre, cette activité dans le PFC est corrélée à la durée des essais (*time to goal*), tandis que l'activité phasique des neurones DAergiques de la VTA est inversement corrélée avec ce paramètre. Ces deux activités interviennent toutes deux précocement dans la séquence, et auraient une influence antagoniste sur la motivation globale des animaux dans la tâche. L'inhibition de l'activité phasique DAergique par optogénétique ayant été réalisée uniquement dans le cadre déterministe, nous n'avons pas pu questionner les relations temporelles avec l'activité prédictive dans le PFC. Pour déterminer ces relations, une

manipulation en cadre probabiliste pourrait être réalisée. Néanmoins, nous serions confrontés à une difficulté d'interprétation : cette manipulation affecterait probablement à la fois l'encodage du résultat du choix précédent (*outcome*), mais aussi du choix futur. Son effet potentiel sur l'activité du PFC impliquerait donc plusieurs variables confondantes. Une autre solution serait d'enregistrer l'activité DAergique de la VTA et du PFC simultanément chez les mêmes animaux, afin de dégager des relations temporelles précises.

Au contraire de ces motifs électrophysiologiques précoces, l'oscillation δ (3-6 Hz) dans le PFC que nous avons relié dans la tâche à un signal d'omission survient plus tardivement dans la séquence d'activité. Cette activité ne corrèle pas avec les choix des animaux, suggérant un rôle double du PFC dans le traitement d'informations à la fois passées et futures.

L'architecture de l'activité neuronale lors d'un essai met en évidence, sur un circuit constitué de la VTA, de l'OFC et du PFC, différents codages (valeur, incertitude, initiation d'action, vigueur) distribués dans le circuit. Les sous-circuits agissent parfois de concert, et parfois en opposition. Cette mesure dynamique de l'activité au sein de différents sous-circuits est rendue possible par la structure de la tâche qui permet un alignement précis des comportements et des activités neurales. Le schéma d'activités que nous mettons en évidence suggère un codage distribué et dynamique avec des calculs réalisées en parallèle (Cisek, 2012; Hunt and Hayden, 2017). Cette vision se démarque, me semble-t-il, des nombreuses théories suggérant que les actions dirigées vers un but peuvent être comprises comme le résultat d'une série de traitements localisés et modulaires, chaque module calculant une variable d'un modèle phénoménologique de comportement (Gershman and Uchida, 2019).

2) Dépendance au contexte

La séquence d'activité précédemment décrite est modulée par le contexte de la tâche. Tout d'abord, il semble que le PFC soit recruté uniquement dans un cadre probabiliste. En effet, dès l'introduction des premières omissions au début du cadre probabiliste, des oscillations δ sont observées, alors qu'elles n'étaient pas présentes en cadre déterministe. Cette activité peut donc être reliée à un signal de détection d'omissions de récompenses. A la fin du cadre probabiliste, l'activité précoce de population du PFC prédit l'incertitude associée à la cible 50% (qui est celle associée à la récompense la plus incertaine), ce qui correspond au rôle prêté au

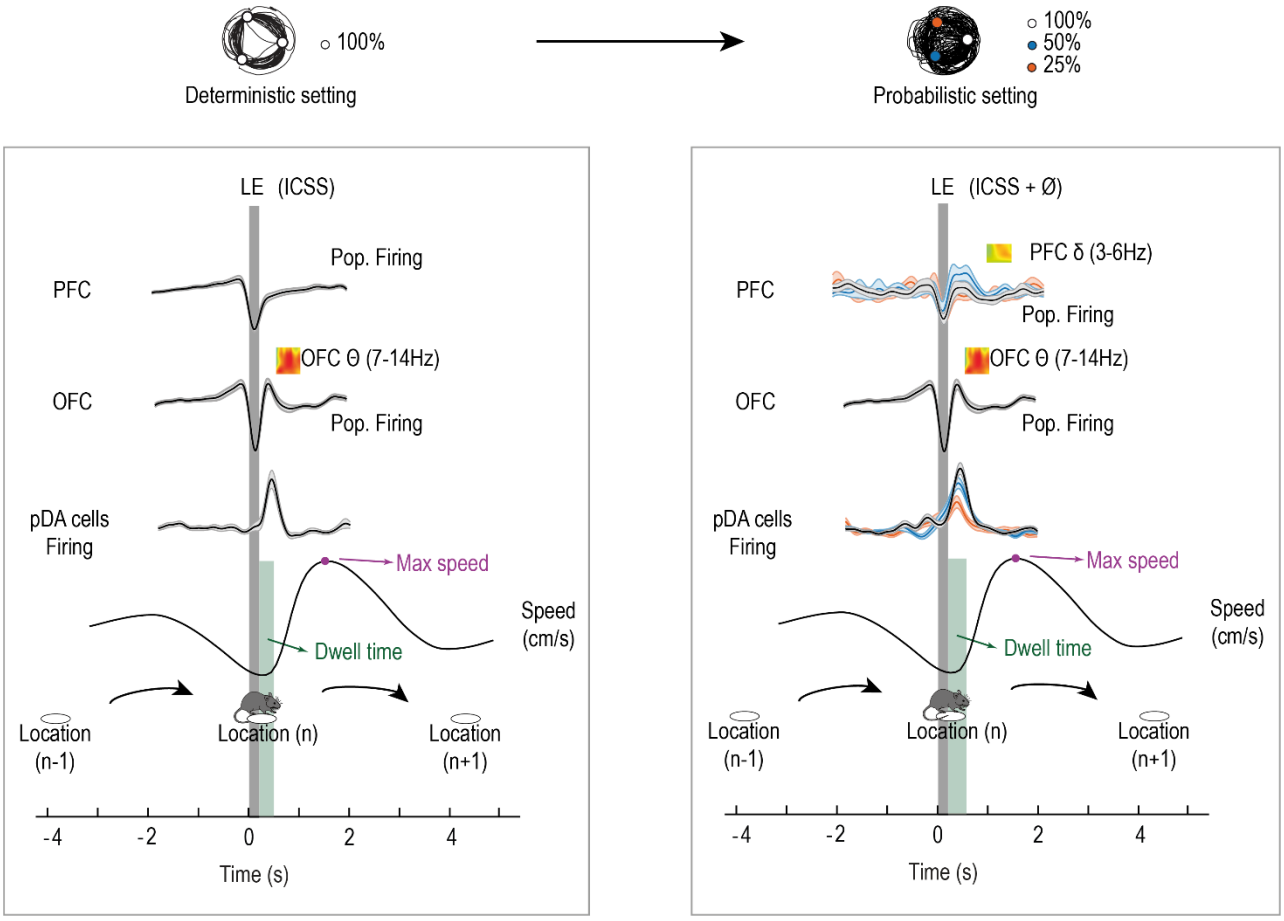
PFC de participer aux décisions sous incertitude (Miller and Wallis, 2013; St. Onge and Floresco, 2010). Par ailleurs, l'activité phasique précoce des neurones DAergiques dépend de la valeur (la probabilité) de la récompense choisie par l'animal.

L'activité du PFC et de la DA dépendent donc toutes deux des paramètres liés à la distribution des récompenses dans la tâche (la probabilité et l'incertitude), c'est-à-dire du contexte de la tâche. Cela correspond en effet à ce à quoi nous pouvions nous attendre d'après ce que rapporte la littérature. En effet, le PFC est impliqué dans le contrôle de changements flexibles de comportements suite à des changements environnementaux (Miller and Wallis, 2013; Rich and Shapiro, 2009; Sharpe et al., 2019). Par ailleurs, certains auteurs rapportent que l'activité prédictive DAergique varie selon que le contexte décisionnel soit un contexte de choix forcé ou libre : dans le premier, elle encoderait la valeur de l'*outcome* du choix, tandis qu'elle encoderait la valeur de l'option disponible de plus haute valeur dans le second (Roesch et al., 2007).

Il est possible que, lors des premières omissions au début du cadre probabiliste, le signal de changement de contingences par le PFC (oscillation δ tardive) et par les neurones DAergiques (RPE négative) enclenche une réorganisation de la séquence d'activités qui amène une représentation différente du contexte environnemental. Nous pourrions tester cette hypothèse en inhibant spécifiquement ces activités lors des premières sessions du contexte probabiliste.

Le contexte (déterministe ou probabiliste) fait émerger une séquence d'activité spécifique (Figure 17). L'activité phasique des neurones DA de la VTA et l'oscillation θ dans l'OFC se mettent en place progressivement au cours de l'apprentissage. L'activité spécifique du PFC semble par contre apparaitre immédiatement en cadre probabiliste, dès les premières omissions. Néanmoins pour qu'une réponse à une omission puisse apparaitre, il faut que la distribution de la récompense soit considérée comme certaine. Il serait donc intéressant de voir si les dynamiques qui se mettent en place en cadre probabiliste sont les mêmes selon que l'animal ait été plongé directement dans un cadre probabiliste ou, comme c'est le cas dans nos résultats, ait d'abord appris le cadre déterministe. Enfin, l'activité phasique DAergique, tout comme l'oscillation θ dans l'OFC, émerge au cours de l'apprentissage de la tâche déterministe. Cependant, ces dynamiques semblent préexister et ne se développent pas à proprement parler. Ainsi, les neurones DAergiques présentent déjà une activité en bouffées de PA (*bursts*) et l'OFC des oscillations θ au début de la tâche et en activité spontanée, et on observe une réorganisation de ces activités à des temps particuliers de la séquence comportementale. La proportion

d'activité phasique dans les neurones DAergiques ou d'oscillations θ dans l'OFC augmente avec les mesures reflétant l'apprentissage de la tâche (augmentation du nombre de récompenses obtenues, diminution de la durée des essais et du temps de résidence, etc.). Or, nous n'observons pas d'augmentation globale de l'activité en fréquence des neurones DAergiques, ni de la puissance globale θ dans l'OFC, qui aurait lié l'émergence de ces dynamiques à une simple augmentation du nombre d'essais. Cela suggère effectivement une réorganisation dynamique et apprise de ces activités. La question est de savoir si cette réorganisation supporte l'apprentissage ou en est la conséquence.



<u>FIGURE 17 :</u> La séquence d'activités mésocorticales dépend du contexte

Représentations schématiques de la séquence d'activité des neurones DAergiques de la VTA, et de l'activité de population ainsi que des oscillations θ et δ dans l'OFC et le PFC, à la fin du cadre déterministe (gauche) ou probabiliste (droite). Les profils de vitesse instantanée, la durée des pauses sur les points (*dwell time*) et l'amplitude des activités présentées ne sont pas précisément à l'échelle, seul un ordre de grandeur temporel relatif est respecté. Notons que ces représentations ne sont pas issues des mêmes animaux. LE : Détection de l'animal sur le point (*location entry*). ICSS : Autostimulation intracrânienne. Ø : Omission d'ICSS.

Ce travail suggère donc qu'une séquence d'activités distribuées, impliquées ou corrélées aux comportements ou aux choix des animaux, est remaniée selon le contexte de la tâche. Cette réorganisation s'opère sans indice externe pour synchroniser ces dynamiques, mais selon une référence interne à l'animal. Les décisions effectuées de manière auto-générée par les animaux reposent donc sur des changements intrinsèques aux représentations de l'animal. Ceci dit, plusieurs questions demeurent.

Les interactions entre neurones DAergiques, l'OFC et le PFC sont-elles issues d'un processus continu qui s'opère selon un drive interne ? Ou bien sont-elles plutôt enclenchées suite à des représentations saillantes émergeant du contexte (par exemple les premières omissions de récompense) ? Si tel est le cas, comment imaginer ce processus en cadre complexe (cf. Article 2) ? Le comportement de l'animal dans cette tâche 3 points est dirigé vers un but ayant un gain espéré qui est représenté par différents encodages organisés en séquence dans la VTA, du PFC et de l'OFC. Mais dans le cadre complexe, si le but est de ne pas répéter des séquences de choix, comment est-il représenté par ces structures ? On peut imaginer que pour s'opposer à la répétition de mêmes séquences de choix, une désynchronisation entre les activités de ces trois structures est requise.

Par ailleurs, dans le cadre complexe, les omissions sont de nature différente du cadre probabiliste. Dans le premier cas, l'absence de récompense dépend de l'historique des choix des souris, tandis que dans le second, elles sont distribuées aléatoirement avec une probabilité fixe en chaque point. On peut alors imaginer, au contraire, des représentations différentes qui initieraient une nouvelle réorganisation dans la séquence d'activité et soutiendrait une autre forme d'exploration dans les choix. Par exemple, il est possible qu'en cadre complexe, la nouvelle séquence d'activités neuronales intervienne en particulier après les omissions, afin de modifier la stratégie de choix en cours (passer de « tourner dans un sens » à « faire un demi-tour »), ce qui permettrait de générer de la complexité.

III) L'exploration dans la prise de décision

Dans ce travail de thèse, nous avons relié différentes activités dans les neurones DAergiques de la VTA, dans l'OFC et le PFC aux paramètres contextuels de la tâche (récompense, probabilité et incertitude) ainsi qu'à différents aspects comportementaux des actions des animaux (initiation et vigueur). Ces relations ont été attribuées sur la base d'approches corrélatives et causales. Nous nous sommes également intéressés aux stratégies de choix des souris dans différents contextes (déterministe, probabiliste et complexe), et notamment à l'exploration. L'exploration est à la fois considérée comme une stratégie de décision permettant d'acquérir de l'information sur son environnement et de réduire l'incertitude, mais aussi comme issu d'un *drive* qui pousse les individus à intégrer de la variabilité dans leurs choix. Dans ce cadre, je vais tenter d'intégrer les différentes analyses électrophysiologiques et celles issues de la modélisation des choix des animaux.

1) La recherche d'incertitude

Notre étude électrophysiologique dans le cadre probabiliste de la tâche révèle un rôle spécifique du PFC face à l'incertitude : des oscillations δ (3-6 Hz) émergent suite aux omissions, et l'activité de population prédit le choix de plus haute incertitude. Ces résultats sont en accord avec les travaux impliquant le PFC dans le contrôle des stratégies de décision, et notamment son rôle prédominant dans l'adaptation à des changements de règles (*switching*) (Cohen et al., 2007; Daw et al., 2005; Rich and Shapiro, 2009). L'inactivation du PFC par infusion locale de muscimol en cadre probabiliste altère les choix des animaux : ils visitent au total moins souvent la cible associée à la probabilité de récompense 100%, et leurs choix se traduisent par une augmentation du paramètre φ reflétant le bonus d'incertitude (Figure 18). L'augmentation de φ alors que les deux autres paramètres β et κ ne sont pas affectés reflète une augmentation de recherche d'incertitude.

Autrement dit, l'inactivation du PFC accroit chez les souris l'attrait des récompenses incertaines. La recherche d'incertitude a été reliée à de l'exploration dirigée (Cohen et al., 2007). Dans cette expérience, nous n'observons pas particulièrement plus de choix exploratoires dans chaque *gamble,* mais on constate en moyenne une diminution de visite de la cible 100%, c'est-à-dire une diminution d'exploitation. Tandis que l'activité dans le PFC

représente et prédit le point associé à la plus haute incertitude dans la tâche (p = 50%), ce résultat laisse penser que le PFC aurait plutôt tendance à s'opposer à l'exploration. Cette observation est en contradiction avec les études qui l'impliquent directement dans l'exploration et les choix sous incertitude (St. Onge and Floresco, 2010). Notons que certaines études font référence à la notion de risque, et rapprochent le PFC des comportements de recherche du risque. Cependant, lorsque les paradigmes utilisés ne comportent pas de notion de perte ou de punition, mais seulement de distribution probabiliste de récompense, il est selon moi plus légitime de parler d'incertitude que de risque. Nous suggérons que de l'incertitude attendue influence positivement le système motivationnel DAergique, tandis qu'elle a une influence négative sur les systèmes délibératifs. Dans le premier cas, la variabilité de la récompense incite à explorer l'option associée, tandis que dans le second cas la variabilité est déjà estimée et donc considérée comme peu informative, ce qui n'incite pas à choisir cette option.

La DA également a été liée à la recherche d'incertitude (Costa et al., 2014; Naudé et al., 2016). Les souris $\beta2^{-/-}$, présentant une activité DAergique plus faible en fréquence, et une modification de leur activité en bouffées, se comportent différemment de leurs congénères WT en cadre probabiliste (Naudé et al., 2016). Elles présentent une diminution de choix d'exploration qui se traduit par une diminution du bonus d'incertitude φ (Figure 18). Si l'activité phasique des neurones DAergiques est liée à une RPE et à une prédiction de valeur, ces résultats suggèrent que l'activité globale de ces neurones encode également une composante de bonus d'incertitude. Il est alors possible que l'activité phasique DAergique intègre à la fois la valeur et l'incertitude des récompenses (Dabney et al., 2020; Lak et al., 2014), peut-être pas au niveau de chaque cellule mais au niveau de la population neuronale.

Au contraire, dans un article du laboratoire récemment publié (cf. Annexe 1) auquel j'ai participé durant ma thèse, des souris WT ont été soumises à une exposition chronique de nicotine afin d'élever l'activité globale (en fréquence et en *burst*) des neurones DAergique de la VTA. Cette exposition a modifié les choix des animaux en cadre probabiliste vers plus d'exploitation, et a conduit à une augmentation du paramètre β reflétant la sensibilité à la valeur. De même, l'augmentation globale de l'activité des neurones DAergiques de la VTA par optogénétique en utilisant la CatCh (ChR2) durant la tâche suffit à modifier les choix des animaux et à augmenter le paramètre β. Ces résultats (Figure 18) indiquent plutôt que l'activité DAergique est liée à une différence de valeur et ne permettent pas de l'impliquer dans la recherche d'incertitude. Il faut néanmoins souligner que ces expériences concernent une manipulation globale de l'activité du système DAergique et pas spécifiquement de l'activité phasique.

Globalement, l'activité DAergique est un substrat complexe d'exploration car a été reliée à deux stratégies d'exploration dirigée : l'une est dirigée par la valeur des options (cf. Annexe 1), l'autre par un bonus d'incertitude qui motive à explorer (Naudé et al., 2016). Dans ce travail de thèse (Article 1), nous avons mis en évidence que le PFC a un rôle de frein vis-à-vis de la recherche d'incertitude. Cela amène à se questionner sur sa participation dans les expériences mentionnées précédemment. Est-il plus ou moins actif en cadre probabiliste lorsque l'activité DAergique globale est modifiée ? L'élévation de l'activité DAergique par stimulation optogénétique de la CatCh (Annexe 1) modifie-t-elle les représentations d'incertitude ou de *mismatch* dans le PFC ? On peut imaginer, en accord avec la littérature (Durstewitz and Seamans, 2008; Seamans and Yang, 2004), que plus de DA stabiliserait plus facilement ces représentations dans le PFC, ce qui, en accord avec les résultats de l'Article 1, aboutirait à une tendance à éviter les choix associés à l'incertitude. Alors, le niveau d'activité DAergique étant surélevé, les signaux associés aux RPE deviendraient plus prépondérants dans la politique de choix, ce qui se traduirait par des stratégies plus sensibles aux valeurs, et donc à un β plus élevé. La DA et le PFC pourraient donc avoir des rôles antagonistes complexes dans les comportements d'exploration et le contrôle de la balance exploration / exploitation. Ils pourraient agir plus ou moins en synergie selon le contexte et orienter ainsi la politique des animaux vers plus d'exploitation ou plus d'exploration.

2) La composante motivationnelle de l'exploration

Les rôles complémentaires des neurones DAergiques de la VTA et du PFC dans le contrôle de l'exploration dans les choix posent également la question de la motivation. En effet, on considère que le bonus d'incertitude φ oriente les politiques de choix vers plus d'exploration en augmentant la valeur des options incertaines (cf. Chapitre 4 – III) – 1), équation (2)). Néanmoins, les animaux choisissent d'explorer même lorsqu'ils se trouvent dans des états où l'on suppose qu'ils ont de bonnes estimations des paramètres de l'environnement. En effet, nous considérons, sur la base de leurs choix et trajectoires, qu'après avoir été soumis pendant 10 jours de suite aux mêmes probabilités de distribution de récompenses dans le cadre probabiliste de la tâche, les animaux ont atteint un état stable et ont appris, ou approximé, les valeurs associées à chaque cible. Dans cet état stable, l'exploration en particulier du point 50%

persiste dans le pari G_{25}. Cette exploration résiduelle peut également être reliée à un *drive* motivationnel.

Dans ce travail de thèse, le PFC et la DA ont également été corrélée ou causalement impliquées dans la motivation des animaux dans la tâche. L'activité phasique DAergique a été causalement liée à l'initiation des actions (*dwell time*), leur vigueur (temps de vitesse maximale) et leur durée (*time to goal*). Le PFC quant à lui a été causalement lié à la motivation globale (*time to goal* et nombre de récompense). Le PFC et la DA, qui sont deux acteurs des politiques d'exploration dirigée, participent également à l'état de motivation général des actions auto-générées dans la tâche. L'activité phasique DAergique est inversement corrélée à la durée des essais, tandis que la relation est positive avec l'activité de population du PFC. Par ailleurs, l'augmentation générale de l'activité des neurones DAergiques via une exposition chronique à la nicotine, ou bien via la stimulation optogénétique de CatCh, aboutit également à une diminution de la durée des essais dans la tâche (Annexe 1). Dans ce cas, on peut se demander si l'activité DAergique ne représenterait pas un signal motivationnel d'exploitation plutôt qu'un signal de valeur.

Dans le cadre complexe de la tâche, (cf. Article 2), nous avons montré que les souris étaient capables d'apprendre à générer des séquences de choix de moins en moins répétitives et ce, en utilisant une politique de choix davantage basée sur de l'aléatoire sans mémoire que basée sur la mémoire des choix précédents. Cela se traduit par une politique de choix marqué par un paramètre β de faible valeur (Figure 18). Ces résultats sont en faveur de l'existence d'un *drive* d'exploration qui favoriserait l'introduction de variabilité dans la politique de choix, ou de la possibilité de réduire un drive d'exploitation (si l'on considère que, sans informations, le système choisit au hasard et que l'apprentissage des options pose des contraintes et biaise le système). Cette stratégie d'exploration stochastique signifierait que les décisions des animaux sont effectuées en partie avec une dimension aléatoire interne et non conséquente à une complexité dans les stimuli, comme nos résultats le suggèrent (cf. Article 2). En effet, dans le cadre complexe de la tâche, la distribution des récompenses n'est pas stochastique, elle est déterministe dans le sens ou si un animal reproduit une séquence, le résultat de cette séquence de choix sera toujours le même en termes de récompense. Les animaux ont donc la possibilité, en théorie, d'extraire un certain nombre de récurrences pour augmenter leur gain. Les modèles que nous avons utilisés pour reproduire les choix des souris indiquent que les choix ne sont pas basés sur la mémoire des visites précédentes mais plutôt sur de l'exploration aléatoire. Cela

suggère aussi que les décisions ne sont pas orientées par les contingences de la tâche mais plutôt selon une référence interne à l'animal. La question qui subsiste est de savoir dans quelle mesure cette stratégie de choix basée sur la « non répétition » s'oppose à l'idée de décisions orientées vers un but.

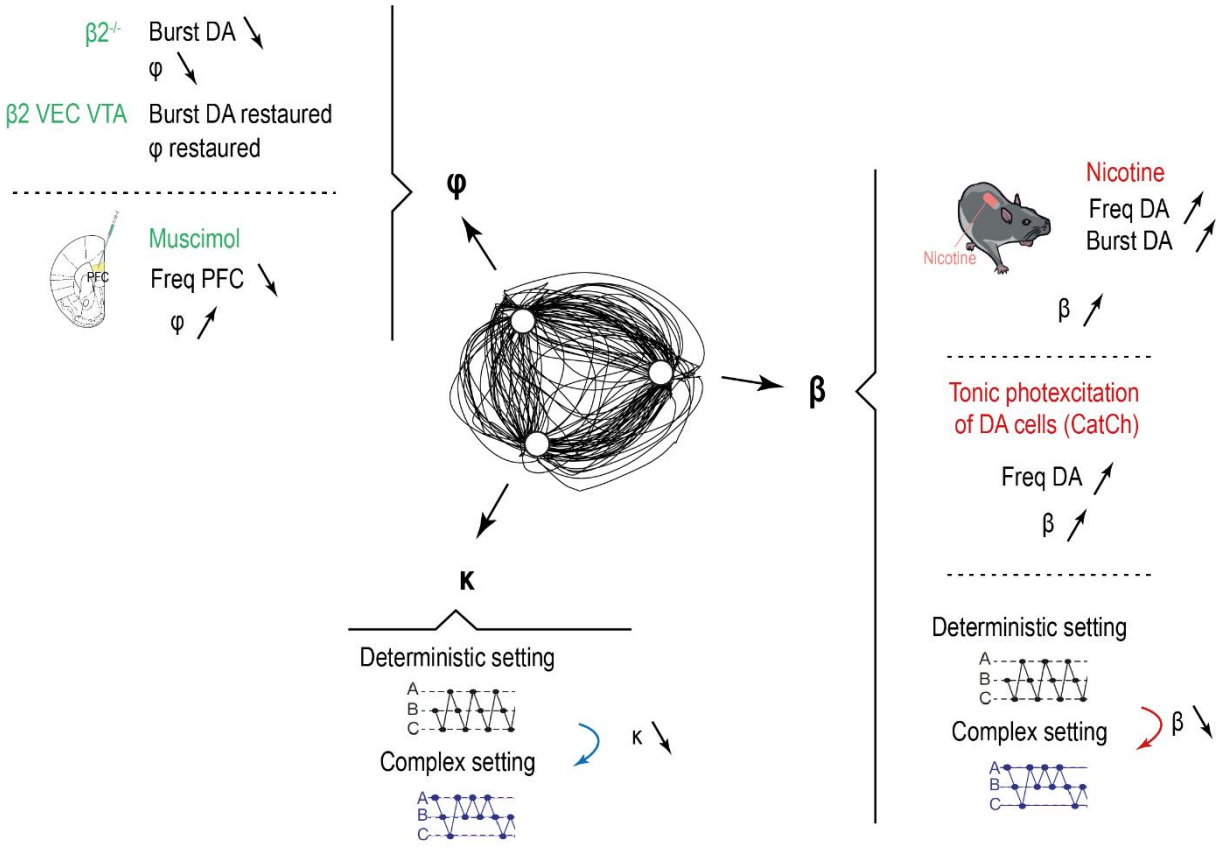


FIGURE 18 : Différents substrats du contrôle des décisions dans la tâche

Schéma-bilan résumant les liens mis en évidence au cours de ma thèse entre les 3 paramètres du modèle de comportement de choix et l'activité du système mésocortical dans les différentes variantes de la tâche 3 points. Dans le modèle, φ est un bonus à l'incertitude qui est intégré dans le calcul de la valeur subjective associée à une cible et oriente donc l'exploration vers l'option la plus incertaine. Des travaux précédents de l'équipe (Naudé et al., 2016) ont montré dans le cadre probabiliste que la sous-unité β2 du récepteur nicotinique dans la VTA était nécessaire à la fois à l'existence d'un tel bonus à l'incertitude mais aussi à l'activité en bouffées des neurones DAergiques de la VTA chez l'animal anesthésié. Mes travaux ont également montré que l'inactivation du PFC avait tendance à augmenter le bonus à l'incertitude φ (cf. Article 1). Le paramètre β représente la sensibilité à la différence de valeurs entre deux options. Un β élevé signifie qu'un individu va accorder une forte préférence aux options les plus récompensantes et moins explorer les autres. L'élévation de l'activité DAergique de la VTA, soit par une exposition chronique à la nicotine, soit par stimulation optogénétique, augmente β et pousse les animaux vers un profil exploitant (cf. Annexe 1). A l'inverse, la transition de la règle déterministe vers la règle de complexité diminue le β ainsi que le coût du demi-tour κ : l'animal sort d'une stratégie d'exploitation de schémas répétitifs et explore davantage en générant des séquences pseudo-aléatoires (cf. Article 2).

Globalement, les différentes expériences menées au cours de ma thèse concluent que le niveau global d'excitabilité du système DAergique est impliqué dans la sensibilité à la valeur (β) et le contrôle de la balance entre l'exploitation et l'exploration. Il est également possible qu'il soit impliqué dans des stratégies de choix plus aléatoires. La synchronisation des PA des neurones DAergiques dans le *burst* est, quant à elle, davantage impliquée dans l'encodage de bonus de valeur notamment le bonus à l'incertitude (φ). Il existe probablement un continuum entre ces deux dimensions. Nous pouvons supposer, au regard de ces travaux, que l'OFC et le PFC orientent sur différentes échelles (encodage des valeurs, détection d'omissions, prédictions d'incertitude) l'intégration de ces deux modes d'activité et participent activement à la génération de décisions adaptées au contexte.

Annexe 1 – Dongelmans et al. 2020

Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in a foraging task. Dongelmans M*, Durand-de Cuttoli R*, Nguyen C*, Come M*, Durante E K, Lemoine D, Brito R, Ahmed Yahia T, Mondoloni S, Didienne S, **Bousseyrol E**, Hannesse B, Reynolds L, Torquet N, Dalkara D, Marti F, Mourot A, Naudé J, Faure P. Nature Communications (2021).

Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration.

Malou Dongelmans[1#], Romain Durand-de Cuttoli[1,2#], Claire Nguyen[1#], Maxime Come[1,3#], Etienne K. Duranté[1], Damien Lemoine[1], Raphael Brito[1], Tarek Ahmed Yahia[1], Sarah Mondoloni[1], Steve Didienne[1,3], Elise Bousseyrol[1,3], Bernadette Hannesse[1], Lauren M. Reynolds[1,3], Nicolas Torquet[1], Deniz Dalkara[4], Fabio Marti[1,3], Alexandre Mourot[1,3], Jérémie Naudé[1,3], Philippe Faure[1,3*]

Affiliations

[1] Sorbonne Université, INSERM, CNRS, Neuroscience Paris Seine - Institut de Biologie Paris Seine (NPS - IBPS), 75005 Paris, France.

[2] Present address: Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3] Brain Plasticity Unit, CNRS, ESPCI Paris, PSL Research University, 75005 Paris, France

[4] Sorbonne Université, INSERM, CNRS, Institut de la Vision, Paris, France.

[#] These authors contributed equally:

[*] Correspondence to phfaure@gmail.com

Abstract

Long-term exposure to nicotine alters brain circuits and induces profound changes in decision-making strategies, affecting behaviors both related and unrelated to drug seeking and consumption. Using an intracranial self-stimulation reward-based foraging task, we investigated the impact of chronic nicotine on the trade-off between exploitation and exploration, and the role of midbrain dopamine neuron activity in decision-making unrelated to nicotine-seeking. Model-based and archetypal analysis revealed substantial inter-individual variability in decision-making strategies, with mice passively exposed to nicotine shifting toward a more exploitative profile compared to non-exposed animals. We then mimicked the effect of chronic nicotine on the tonic activity of dopamine neurons using optogenetics, and found that photo-stimulated mice had a behavioral phenotype similar to that of mice exposed to nicotine. Our results reveal a key role of tonic midbrain dopamine in the exploration/exploitation trade-off and highlight a potential mechanism by which nicotine affects the exploration/exploitation balance and decision-making.

Introduction

Nicotine is the primary reinforcing component driving tobacco addiction [1–3]. Like most addictive substances, nicotine is hypothesized to perpetuate addiction through alterations in dopamine (DA) signaling and plasticity in the mesocorticolimbic pathway [4]. Repeated activation of ventral tegmental area (VTA) DA neurons by nicotine not only leads to reinforcement but also to craving and lack of self-control over intake [5]. Concurrently, chronic exposure to nicotine exerts numerous effects on brain and circuits, affecting personality traits and behaviors that extend beyond drug-seeking or consumption, [6,7] such as changes to emotional state or levels of stress [8,9] and anxiety [10]. Chronic nicotine exposure also impacts various components of decision-making processes, such as impulsivity[11,12] or exploratory behaviors [13,14] which may contribute to the persistence of drug consumption by promoting relapse and susceptibility to other addictions [15]. However, directly linking the cellular effects of nicotine to modifications of decision-making has been elusive. Understanding the molecular and circuit-level mechanisms of nicotine on decision making is needed to decipher its multifaceted effects. Here we take advantage of a decision-making framework in a rodent model to address the impact of chronic nicotine exposure on VTA DA neuron activity and decision-making parameters.

Among the components of decision making, the explore/exploit tradeoff is of particular interest. Exploitation refers to choosing the option that seems, based on the history of rewards, the optimal choice. However, when faced with a choice between two alternatives, one with low and one with high probability of reward, animals don't purely exploit, they also choose the less likely rewarded option a significant portion of the time. The origin of seemingly suboptimal choices, however, remains poorly understood. It has been interpreted in different studies as noise, error, risk seeking, irrational belief or exploration [7,16–19]. In the context of exploration, choosing an option with less likelihood of immediate reward is essential to gather information about unknown or uncertain outcomes in a changing environment. Exploration is thus central to the emergence and organization of behaviors [20], naturally resulting in the acquisition of new information crucial for learning and optimizing behavioral strategies [7,17]. Furthermore, as both exploitation of reward knowledge and exploration of reward options constitute important, yet opposing, adaptive processes, determining the exact trade-off between exploration and exploitation is key to decision-making. This trade-off is ubiquitous across species and pervades a number of altered behaviors under specific psychiatric condition such as addiction [6,7]. It is thus an ecologically valid tool for translational research and for dissecting the link between the impact of the drug at the molecular and circuit levels and a relevant behavior. In the context of addiction, a modification of this trade off will impact the global equilibrium of decisions between drug and non-drug rewards. Determining whether chronic nicotine exposure alters such exploration-exploitation trade-off is thus fundamental to help understand modifications of individual traits associated with continued nicotine consumption.

Altered DA function is a promising candidate to link chronic nicotine exposure to changes in decision making behavior. This neuromodulator, which is at the crossroads of motivation, learning and decision-making, can be hijacked, in the context of addiction, by most drugs of abuse [21–23]. Changes in the spontaneous tonic firing of VTA DA neurons, as a consequence of repetitive drug-use, can indeed alter the subjective value assigned to available rewards [21], as well as the motivational salience of the drug or of drug-predicting cues [24], influencing

decisions about which reward to pursue [25]. Tonic DA can scale the performance of a learned behavior [26], the incentive value associated with environmental stimuli [27], or signal the average reward [28]. In the exploration/exploitation framework, the role of tonic DA remains debated. The effect of DA manipulation on the exploration/exploitation balance is convincing but varies depending on the task [29–31]. Increasing tonic striatal DA release has been suggested to either increase [29] or decrease [31] the level of exploration. Decreasing tonic striatal DA has also been suggested to increase exploration [32]. Hence, drug-induced alterations of DA transmission may modify behavioral choices, either positively or negatively depending on the environment and the specific type of DA manipulation.

Using an intracranial self-stimulation (ICSS) reward-based foraging task for mice, we have shown that decisions in this task are modulated by the cholinergic neurotransmission of the VTA, with a particular role of nicotinic acetylcholine receptors in expected uncertainty driven exploration [33]. Here we demonstrated that chronic nicotine exposure increases the tonic activity of VTA DA neurons and reduces exploration to favor exploitation, with mice focusing on the most valuable options at the expense of information gathering. Increasing the tonic activity of VTA DA neurons using optogenetics is sufficient to mimic the behavioral bias (or exploratory decrease) induced by nicotine, suggesting that the DA control of the exploration/exploitation balance is altered by long-term nicotine exposure.

Results

Motor cost, probability, and uncertainty associated with reward shaped mouse choices

To assess choice behavior in an uncertain environment, we used a multi-armed ICSS bandit task for mice where specific locations, hereafter called targets, were associated with brain stimulation rewards delivered to the medial forebrain bundle (MFB) (Figure 1a, Supplementary Figure 1) [19,33]. The task takes place in a circular open-field (interior diameter = 80 cm), with three explicitly marked targets forming the apices of a triangle (Figure 1b). Passing over each target results in the delivery of a rewarding intra-cranial electrical stimulation. Mice can not receive two consecutive stimulations from the same target, and thus learn to forage from one to another to continue receiving stimulations (Figure 1b Left). During the training period (5-min daily sessions), hereafter called the deterministic setting (DS, Figure 1c Left), every visit to a target was reinforced by a stimulation reward (reward probability p = 100% at each location, p_{100}). At the end of the deterministic setting, mice were confronted with a probabilistic setting (PS, Figure 1c Right) where each target was now associated with a different probability of stimulation delivery (p = 100%, 50% and 25%, Figure 1c Right). As previously shown [33], both the expected reward probabilities and uncertainties associated with the different targets in the probabilistic setting induced a marked change in the behavioral pattern compared to the deterministic setting. Trajectories at the end of the deterministic setting were stereotyped, almost circular, with a low probability of directional changes (i.e. returning to the previous target, Figure 1d) due to an associated motor cost [19]. In contrast, in the probabilistic setting mice distributed their choices differently and increased their probability of directional changes, indicating an adaptation from the circular strategy (Figure 1d). Directional changes in the probabilistic setting were not random: rather, they allowed animals to focus on specific targets. Indeed, compared to the deterministic setting where mice visited the three targets with

a uniform distribution, in the probabilistic setting mice visited more often the targets associated with the highest reward probabilities (i.e. p_{100} and p_{50}, Figure 1e). While a purely exploitative strategy would consist solely in an alternation of visits between p_{100} and p_{50}, we found that mice continued to visit all three points, prompting us to further investigate the exploration/exploitation tradeoff in their choices. However, the global repartition of visits does not directly measure choices. Indeed, since mice could not receive two consecutive rewards from the same target, the repartition of visits on the three locations resulted from a sequence of binary choices (Figure 1f) in three gambles (G_{100}, G_{25}, G_{50}) between two respective payoffs (here, G_{100} = {p_{50} versus p_{25}}, G_{25} = {p_{100} versus p_{50}}, G_{50} = {p_{100} versus p_{25}}). Hence, we analyzed the sequence of choice statistics using a transition function, allowing us to investigate free choices in each gamble independently, and the trade-off between exploitation and exploration. When faced with a choice between two alternatives, exploitation corresponds to choosing the option for which the animal assigns the highest value, while exploration corresponds to choosing the less valued alternative. Animals purely exploiting would always choose the high probability option, but we found that mice chose the less likely rewarded option a significant portion of the time, consistent with balancing exploitation and exploration in their choice behavior. For G_{100} and G_{50}, mice chose the optimal location (i.e., the one associated with the highest probability of reward) more than 50% of the time. However, for G_{25} (i.e. the free, binary choice between p_{100} and p_{50} when the animal is on p25) the probability to choose p_{100} over p_{50} was not different from a random choice (Figure 1f), which has been interpreted as indicating that mice assign a positive motivational value to expected uncertainty, which is maximal at p_{50} [33]. Overall mice biased their choices depending on both the probability and the uncertainty of reward delivery. Behavior in the task was therefore the result of a balance between exploratory and exploitative choices.

Nicotine exposure biased choices toward the most valuable options

We aimed to investigate the effects of chronic nicotine exposure on decision-making behavior and on the balance between exploration and exploitation. To do so, we implanted osmotic minipumps subcutaneously to expose mice to continuous nicotine (Nic, 10mg/kg/day) or saline (Sal) for 3 weeks and then compared their behavior in the probabilistic setting of the ICSS task (Figure 2a). Because nicotine induces long lasting adaptations in the midbrain DA system [34], and because VTA DA neurons have been associated with decision-making under uncertainty [22,33], we first analyzed the spontaneous tonic activity of VTA DA cells in anesthetized mice. We recorded from mice chronically exposed to either saline or nicotine via minipump, and that had performed the behavioral task ("ICSS", at the end of probabilistic setting), or were behaviorally naïve. DA neuron firing was analyzed with respect to the average firing frequency and the percentage of spikes within bursts. As previously reported [9,35], chronic exposure to nicotine increased the tonic activity of DA neurons, both in terms of firing frequency and bursting activity, when compared to mice implanted with a saline minipump (Figure 2b). Furthermore, mice exposed to the ICSS task exhibited an increase in firing frequency, but no change in bursting activity when compared to mice that were not stimulated (Figure 2b).

We then analyzed the behavior of mice in the ICSS task. Overall, we did not see any behavioral difference between mice implanted with a saline minipump (n=23) and the non-implanted mice (n=32) analyzed in Figure 1

(Supplementary Figure 2 a-c). Therefore, these two groups were pooled and henceforth referred to as control (Ctl, n=55). Trajectories at the end of the deterministic setting were stereotyped, almost circular, in both control and nicotine-treated mice. Both groups distributed their visits equally over the three locations (Figure 2c) and their respective probabilities of directional changes were equal (Δ = -2.7%, Figure 2d). However, the total number of rewards was higher for nicotine-treated than for control mice (Δ = 26, Figure 2e), as a consequence of the decrease in the mean time-to-goal (i.e., the time necessary to go from one target to the next) in nicotine-treated mice (Δ = 0.83 s, Figure 2f). When mice were placed in a classical open field (without ICSS), a greater velocity was observed in mice exposed to nicotine, yet only at the beginning of the session (Supplementary Figure 2d). This result suggests that the increased speed observed in the ICSS task for nicotine-treated mice may arise from the combined effects of nicotine exposure and the stimulation rewards.

Clear differences in the behavior of nicotine- and saline-exposed mice were observed in the probabilistic setting. Both groups distributed their choices depending on the probability to receive a reward, but with markedly different strategies. Notably, while control mice visited significantly p_{25}, nicotine-treated mice instead focused on visiting the two most rewarded options (i.e., p_{50} and p_{100}, Figure 2g, Δ_{25} = -5%, Δ_{50} = 2.7 %, Δ_{100} = 2.3%), which was associated with an increase in the percentage of directional changes (Δ = 11%, Figure 2h). These alterations in overall repartition resulted from changes in successive binary choices, with an increase of the optimal choice in gamble G_{100} (Figure 2i, Δ = 10%) for nicotine-treated mice compared to control mice. We also observed an increase in the total number of obtained rewards (Δ = 17.9, p = 0.002) and in the percentage of success (number of rewards divided by the number of trials, Δ = 2 %, p = 0.02) in nicotine-treated mice compared to control mice. Finally, the comparison of mean time-to-goal between the two groups (Δ = -1.1 sec, Figure 2j) indicates again an increased velocity in nicotine-treated mice, as was already observed in the deterministic setting. This increase in speed in the probabilistic setting was not associated with a decrease in the number of directional changes made by nicotine-treated mice, suggesting that animals did not enter an automatic circular mode, disengaged from actual choices, but instead remained in a deliberative process. Altogether, these results indicate that chronic nicotine modifies the decision strategy of mice by biasing choices toward the seemingly most valuable options.

In the probabilistic setting, adopting a purely exploitative strategy to maximize the success rate would require mice to choose the alternative with the highest probability of reward in each gamble, leading to a sequence of choices with solely the alternation of visits between p_{100} and p_{50}. Both control and nicotine-treated groups clearly deviated from this strategy of pure exploitation, although nicotine-treated mice were more exploitative on average. Yet population analyses (i.e. averaging over groups of animals) classically do not reflect the wide range of distinct behaviors and strategies that can be adopted by individuals. We therefore further analyzed our behavioral data, with the aim of revealing individual profiles and their adaptation under nicotine exposure.

Idiosyncrasy in choice behavior suggests individual strategies

Visual inspection of individual trajectories revealed that in the probabilistic setting, some mice retained a circular strategy (with either an ascending (p_{25} - p_{50} - p_{100}) or descending (p_{100} - p_{50} - p_{25}) order) while others had what we hereafter call a gain-optimizing (GO) strategy, alternating between targets associated with the highest

reward probabilities (p_{100} and p_{50} - Figure 3a, lower left). By a gain-optimizing strategy, we mean a very basic definition of optimality based only on maximizing the number of rewards, but which does not take into account the potential advantage of exploration. Theoretically, always choosing the most valuable option would lead to an average success rate of 75% (Figure 3a, lower right) while a purely circular strategy would lead to an average estimate of 58.3% success rate (Figure 3a, upper right). Accordingly, the percentage of directional changes was correlated with the success rate (Figure 3b, for control and nicotine-treated mice). Progressively adding directional changes between the p_{50} and p_{100} targets to the circular pattern theoretically result, in this plot, in a displacement along the line that connects the theoretical points of the circular strategy (0% directional change, 58.3% success) to the gain-optimizing strategy (100% directional changes, 75% success) (Figure 3b, red line). We found experimentally that the slope ($s = 17.1 \pm 1.5$, black line, Figure 3b) of the correlation between percent of directional change and success rate was almost parallel to the theoretical line from circular to gain-optimizing strategies (S_{th}= 16.7, red line, Figure 3b), indicating that most of the directional changes were not random, but consisted in back-and-forth sequences between the p_{50} and p_{100} targets.

Inter-individual variabilities of choice patterns in mice were not the consequence of random variations or different stages in the same learning process. Rather, they were robust in each animal throughout the task, indicating differences in individual strategies. This is suggested first by the overall stability of the behaviors as indicated by the convergence to a plateau at sessions 8-10 (Supplementary Figure 4a), and by the absence of any positive correlations between decision-making parameters and session number after the first 5 sessions (Supplementary Figure 4b). Furthermore, to test whether the variabilities in behavior were robust for each individual from trial to trial, we compared the percentage of directional changes for two consecutive sessions for each animal of the control group. Directional changes showed a strong positive correlation from one session to the next (Figure 3c), suggesting a strong consistency in individual behaviors. This observation was generalized by demonstrating that intra-individual variations are lower than the inter-individual variations (Supplementary Figure 4c).

Having established that inter-individual variations in the probabilistic setting performance arise from the strategies each mouse adopts within the task, we next aimed to characterize individual behaviors of all mice (both control and nicotine-treated groups, i.e n=82) in the task. For that purpose, we used a seven-dimensional dataset based on the statistics of i) the directional changes, ii) the target distributions and iii) the three gambles (see data Figure 1d-f) followed by archetypal analysis [36,37]. While principal-component analysis methods have been classically used to split high-dimensional data sets into clusters by aggregating individual data onto typical observations (the cluster centers), archetypal analysis depicts individual behavior instead as a continuum within an "archetypal landscape" defined by extreme strategies, the archetypes. Individual data points are represented as linear combinations of extrema (vertex corresponding to "archetypal strategies") of the dataset. The seven-dimensional dataset was used to identify three archetypal phenotypes. The three archetypes and their characteristics (Figure 3d) differentiated mice exhibiting a GO strategy (i.e. focusing on p_{50} and p_{100}), Figure 3a, in grey)), from mice with circular patterns (equal distribution between the three targets, Figure 3a), which either turned in a descending manner (labelled Des, in blue, sequence p_{100} - p_{50} - p_{25} associated with high G_{100} and G_{25} but low G_{50}) or an ascending manner (labeled Asc, sequence p_{25} - p_{50} - p_{100} associated with low G_{100} and G_{25} but high G_{50}).

The individual behavior of each of the 82 mice could be defined as a weighted combination of these three extremes in a ternary plot (Figure 3e). An animal's behavior in this ternary plot is defined by three coordinates (a,b,c) that sum to 1 and that depict its relative archetypal composition. Therefore, these coefficients (a,b,c) could be used to assign each individual to its nearest archetype based on its behavioral profile (Figure 3e, left). This assignment revealed that 23.2 % of the mice were closer to the GO archetype (grey), while the remaining mice were evenly distributed between the Des (39%, blue) and Asc archetypes (37.8%, green) (Figure 3e, right). To analyze the effect of chronic nicotine, we split the control and nicotine-treated mice, and showed that these two groups are distributed differently in the archetypal space as indicated by a modification of i) the distribution of the archetype's assignments (Figure 3f) and of ii) the archetypal composition (Figure 3g). Overall, chronic nicotine exposure produced an apparent displacement of the population further from Asc and Des apices and closer to the GO apex, thus it favored the emergence of the more exploitative, and thus less explorative, GO phenotype.

Nicotine modified decision parameters associated with exploration

To quantitatively describe the effects of nicotine on the decision processes underlying steady-state choice behavior in mice, we modeled individual data using a softmax model of decision-making. In this model, the probability of choosing target A over B depends on the difference between their expected values, here the probability *p* of reward delivery associated with each target (as the stimulation magnitudes were the same for all targets), and the "inverse temperature" parameter β which represents the sensitivity to the difference of values (ΔV). A small β favors exploration (the proportion of respective choices is less sensitive to ΔV, with a null β meaning all options have nearly the same probability to be selected, independently of their respective value), while a large β indicates exploitation (high sensitivity to ΔV, with an infinite β meaning that options associated with higher reward probabilities are always selected). β can thus be considered as a proxy to measure the exploration/exploitation tradeoff. "Choosing the highest rewarded option" and "exploring less" are therefore equivalent in this exploit/explore framework. This model was adapted to account for the behavior of mice in the probabilistic setting as follows: first, decisions were biased towards actions with the most uncertain consequences, by assigning a bonus value φ to the expected uncertainties, i.e. the variance $p(1-p)$ associated with each location [33]. This allowed us to explain the atypically low probability of choosing p_{100} over p_{50} in G_{25} (Figure 1f). Second, to account for the circular bias observed in both deterministic and probabilistic setting, we added a motor cost which decreases the value of a target if it requires the animal to perform a directional change [19]. Thus, in this adapted softmax model (Figure 4a and Methods), three latent variables not directly observed but inferred from the model were used: the "exploration/exploitation" parameter β, which was defined as the weighted sum of the expected values (100, 50 or 25 %); expected uncertainty (weighted by parameter φ), and expected motor cost (weighted by parameter κ) of a given target.

We fitted the transition function of each mouse from the control group (n = 55) with this model, and obtained positive β, φ and κ values (Figure 4b left). We then compared the output of this model (labelled M3: $\beta>0$, $\varphi>0$ and $\kappa>0$) with two other models, M1 and M2, where we progressively constrain the parameters. In M1, β and

φ are set to 0, hence choices would be solely driven by motor cost (i.e. a bias against U-turns), which could explain circling behaviors independently of the probabilities associated with reward delivery. In M2 φ is set to 0, which would correspond to animals not taking uncertainty into account. Comparison of the models (Bayesian Information Criterion, Figure 4b right; likelihood ratio test for nested models, Supplementary Figure 5)) indicated that M3 provides the best fit for the data, and suggest that mice used reward probabilities and uncertainty of the reward location to drive their choices.

The robustness of the model was then assessed by generating sequences of choices (n = 2000 model choices) for n = 55 mice with their respective model parameters (Figure 4c, see also Supplementary Figure 5). The model accurately reproduced the mean distribution of targets (Figure 4c, Left), the proportion of directional changes (Figure 4c, Middle) and the choice transition function (Figure 4c, Right). Individual transition functions from nicotine-treated mice (n=27) were then fitted by the same model. When compared with the model parameters of control mice, nicotine exposure increased the value sensitivity parameter β, but did not affect the cost of directional changes (κ parameter), nor the uncertainty bonus φ (Figure 4d). We thus asked whether recapitulating these effects on decision parameter β would be sufficient to mimic the effect of nicotine. We modeled the choices (N = 2000) using decision-making parameters from the control population (n=55, as in Figure 3b-c) modified by the average difference observed in the β parameter from nicotine-treated mice. We compared the three main behavioral measures altered by nicotine: i) the probability to choose the most valuable option in gamble G_{100} (choosing p_{50} over p_{25}), ii) the percentage of directional changes and iii) the probability to visit p_{25}. By applying an increase in β (derived from nicotine-treated mice) to the control model parameters, the model accurately reproduced the changes observed in decision-making strategy following chronic nicotine exposure for the three measures (Figure 4e). Conversely, by combining a decrease in β (i.e. subtracting the average effect of nicotine from the nicotine-treated model parameters) we were able to simulate the conversion of a nicotine-treated behavioral profile into a control profile. These results thus suggest a specific effect of nicotine on the β parameter.

Finally, we assessed whether archetypal phenotypes of choice data could be derived from differences in decision-making processes measured by the parameters of our model, by evaluating the value of the three parameters (β, φ, κ) depending on the archetypal composition (see methods). Overall, the three archetypes corresponded to different combinations of the β and φ model parameters (Figure 4f), with an almost homogeneous distribution of motor cost κ. The GO (grey) archetype was associated with a high value of β (corresponding to exploitation) and φ, which is consistent with individuals that favor the alternation between locations associated with higher probability (p_{100} and p_{50}). The Des and Asc phenotypes corresponded to strong circular behaviors and thus to a low sensitivity to value (low β value), that unmasks the motor cost (κ). Des and Asc groups differed by their φ value (Δ = 1.012, p = 0.0079), which was related to the directionality of their preferred rotation: a low preference to uncertainty φ corresponds to mice choosing the certain p_{100} reward over the uncertain p_{50} reward, resulting in a tendency for sequence p_{25} -> p_{100} -> p_{50} observed in Des mice (blue). Conversely, a high preference for uncertainty φ is associated with the reverse sequence p_{25} -> p_{50} -> p_{100} observed in Asc mice (green). Overall, the decomposition of the archetypal phenotypes into their underlying decision-making processes illustrate how

distribution of individual decision-making strategies (Asc, Des and GO) in a population could correspond to transitions in the parameter values from the same model. It also allows us to interpret the effects of nicotine as a coordinated increase of β consistent with a deviation towards the GO profile.

Optogenetic DA neuron stimulation recapitulated the effects of nicotine

Nicotine exposure is known to induce modifications in a number of brain areas [38], including an increase in the tonic activity of VTA DA neurons, as we indeed confirmed in this study (Figure 2b). Furthermore, the tonic activity of DA neurons has been proposed to play a role in the balance between exploration and exploitation [29–31]. We thus asked whether directly and acutely modifying the activity of VTA DA neurons is sufficient to alter decision-making behavior within a session and recapitulate the effects of chronic nicotine in our ICSS task. To specifically and bi-directionally manipulate VTA DA neurons, we expressed either an excitatory channelrhodopsin (CatCh [39]) or an inhibitory halorhodopsin variant (Jaws [40]) in DAT^{iCRE} mice using a Cre-dependent viral strategy (Supplementary Figure 5a). We confirmed in patch-clamp recordings that continuous 5ms-light pulses at 8 Hz (470 nm) reliably increased VTA DA neuron activity in CatCh-transduced mice (Figure 5a), while 500ms-light pulses at 0.5 Hz (520 nm) reliably decreased their activity in Jaws-transduced mice (Figure 5b). We specifically test the hypothesis that optogenetic activation of VTA DA neurons will reproduce effect of nicotine on choice behavior, while optogenetic inhibition will produce the opposite effect.

After mice completed both the deterministic and probabilistic setting in the ICSS task, they went through optogenetic sessions maintaining the probabilistic setting rules, with an alternating schedule of two days with photo-stimulation (ON, photo-stimulation started 5 min prior to the start of the daily session and was maintained throughout the 5min session) and without (OFF) (Figure 5c). During the OFF days, mice were connected to the optical fiber patch-cord but did not receive light stimulation. For each pair of ON/OFF experiments, we estimated the effect of the photo-stimulation on the four main behavioral measures that were altered by chronic nicotine (Figure 5d, Supplementary Figure 6d. Optogenetic activation increased directional changes (Figure 5d left) and decreased the probability to visit p_{25} (Figure 5d right), favoring alternations between p_{100} and p_{50}, similar to the effect of nicotine. Opposite effects were observed for these two parameters when the firing rate was reduced in VTA DA cells using Jaws (Figure 5d). Optogenetic activation reduced time to goal without affecting the choice in the gamble G_{100}, while optogenetic inhibition did not significantly affect either of these two parameters (Supplementary Figure 6d). We fitted the transition function of Catch- and Jaws-transduced mice with our decision-making model. As we observed with chronic nicotine, the effects of photo-activating VTA DA neurons on decision-making during the ICSS task could be modeled as an increase of β (Figure 5e). Photo-inhibition of VTA DA neurons, however, did not significantly affect the exploration/exploitation tradeoff parameter β (Figure 5e). For each pair of ON/OFF experiments, we also estimated the effect of the photo-stimulation on our seven main behavioral measures (Figure 1d-f) and β parameter by calculating for a given measure (M) the difference $\Delta=M_{ON}-M_{OFF}$. These differences were compared with the net effect of nicotine obtained for each of these parameters by subtracting the mean control value from the mean effect of nicotine (Supplementary Figure 6e, red). Overall, it can be observed that Nic and

photostimulation produced a similar pattern of effects on our behavioral measure (Supplementary Figure 6e), while inhibition produced an inverse effect.

Finally, by analyzing decision-making behaviors between the stimulated (ON) and non-stimulated (OFF) conditions in the previously identified archetypal space, we revealed that VTA DA neuron activation draws individual phenotypes towards the GO archetype (i.e, increased GO archetypal composition), while VTA DA neuron inhibition draws individuals away from GO (Figure 5f). Thus, altering the firing pattern of VTA DA neurons, by changing both the motor cost and the balance between exploration and exploitation behavior, is sufficient to bias decision behaviors in the ICSS task, as suggested by our simulations (Figure 4). Furthermore, increasing VTA DA neuron firing mimicked the effects of chronic nicotine exposure on decision-making measures, linking the behavioral alterations with the physiological changes to DA neurons we observed in nicotine-treated mice.

Discussion

Understanding how nicotine affects decision-making has been challenging, because two different physiological aspects need to be distinguished [41] : (1) nicotine as a reinforcer that directly activates the dopaminergic system to produce reinforcement and nicotine-seeking, and (2) nicotine as a neuromodulator that alters nicotine-independent decision-making processes by modifying the dynamics and computational properties of cholinoceptive circuits. Here, using a multi-armed ICSS bandit task, we show that mice passively treated with nicotine forage more frequently at locations with the highest probabilities of reward (p_{50} and p_{100}) compared to naive animals, suggesting a bias in the exploration/exploitation tradeoff toward decreased exploration. These behavioral changes were accompanied by modifications in the spontaneous activity of VTA DA neurons. We further showed that inter-individual variations in foraging strategies emerge between mice, despite the fact that they are all males of the same genetic background. This suggests that animals idiosyncratically adapt their behavior in response to task constraints, rather than converging toward a theoretical “optimal” performance consisting only of maximizing reward. Finally, acute optogenetic activation or inactivation of VTA DA neurons shifts task strategy in agreement with results from nicotine-exposed mice and from computational modeling. Together, these results suggest that modifications of dopamine release, notably induced by chronic nicotine exposure, scale the exploration/exploitation tradeoff.

In our task, mice adapted their choices according to the probability of reward delivery, but they also consistently kept visiting the targets associated with lower reward probabilities in all of the gambles, even after extended training. Such a high level of exploratory behavior is potentially attributable to the setup, which is characterized by the delivery of small rewards, serially repeated gambles with short delays between trials, and learning through experience [42]. In the exploration/exploitation framework, the fact that mice continue to visit targets with the lowest reward probability in each of the gambles, despite intensive learning, can reflect i) exploratory noise, generally modeled via decreased value sensitivity (or increased randomness) of β in the softmax model, ii) directed exploration, if one considers that mice continue to explore locations associated with low reward probability to reduce the uncertainty associated with probabilistic omission, and iii) uncertainty-seeking, which is neither explorative nor exploitive, but considers that mice simply attribute a positive value to expected uncertainty. Our analysis also

introduces the idea of qualitative inter-individual variations or idiosyncrasy in choice strategy. Model comparison first suggests that all mice, even those that are far from the gain-optimizing (GO) archetype, used information about reward probabilities and uncertainty, and that variability does not necessarily imply different strategies (i.e different models). Indeed, the inter-individual variations were well described by a single computational model of decision-making that takes into account the exploration/exploitation tradeoff, uncertainty, and motor cost, and by continuous variations in β and φ, two latent variables [43] inferred through the model. Note that sex or strain differences may constitute another layer of variability [44], which we are currently addressing in ongoing experiments. Finally, despite variations in individual choice behaviors, the consequences of nicotine administration were consistent, with a clear effect on the β parameter, and a strategy biased towards the exploitation of the highest reward values.

This interpretation is supported by two findings. Firstly, we could demonstrate that variations in behavior are not due to chance but indicate individual personalities as revealed by the strong correlation in descriptive parameter values between consecutive sessions (Figure 4b). Secondly, it is conceivable that mice eventually achieve the optimal behavior (alternating solely between p50 and p100), albeit at different rates of learning. In this case, a possible interpretation of our data would be that nicotine facilitates learning [45] and speeds up the convergence toward the GO profile. However, this hypothesis is unlikely, and is not supported by our data. Indeed, despite significant adaptations during the transition between deterministic and probabilistic settings, as well as during the first probabilistic sessions (1 to 4), the animals' behavior at the end of the probabilistic setting is stable overall, and the choice behavior close to steady-state. In addition, manipulating the activity of VTA DA neuron using optogenetics rapidly altered the behavioral strategies of the mice, with kinetics that are incompatible with synaptic plasticity or learning. We thus argue that inter-individual variations in task performance, as well as nicotine effects, should not be interpreted as differences in learning processes or knowledge of the environment, but rather as differences in using the knowledge acquired about the statistical structure of the environment (quantified by variations in β and φ) to develop their strategy within the task.

The increase of β reflects an amplified exploitative behavior, an effect that has been previously linked to enhanced tonic DA activity, which is hypothesized to modulate the bias towards optimal choices [29–31]. In this study, we demonstrate a direct link between the tonic activity of DA neurons and exploitation using electrophysiological and optogenetic approaches. The multi-armed ICSS bandit task enables, through a clear distinction between action selection (choices) and action execution (time to goal), to identify the modified components of value-based decision-making in relation to tonic DA. We explicitly demonstrate an increase in value sensitivity due to nicotine-induced alterations in tonic DA activity. Previous ICSS studies have observed that chronic exposure to drugs sensitizes the brain reward system, and in doing so lowers the stimulation threshold (expressed as a current intensity or stimulation frequency) [46] required for ICSS [47]. Here we expand these results by quantifying the effects of such increased value sensitivity on choices between ICSS-mediated rewarding locations, and further identifying a causal link between these behavioral modifications and increased tonic activity of VTA DA neurons.

In the context of dopamine neuron physiology, activity varies in frequency and in degree of burst firing. Bursting has been defined as successive action potentials separated by less than 80ms [48], occurring on top of a regular “pacemaking” firing activity. Spontaneous activity (regular spiking and bursting) is associated with the

neuromodulatory function of dopamine and its ability to shift ongoing dynamics of target structures. In this context, bursting is not necessarily locked to any behaviorally relevant or salient event. By contrast, phasic activity is related to event-locked increase in firing [22] , which can typically be observed as a synchronous increase in firing rate in a population of neurons, but is not necessarily composed of bursts of action potentials (i.e. it can be single spikes but time-locked to an event during successive trials). Dopamine phasic activity modulates synaptic plasticity [49] and is critical for learning the value of stimulus or actions [50]. The observed increase in VTA DA neuron activity (both in bursts and in firing rate) after nicotine exposure suggests that dopaminergic tone is modified in nicotine-exposed animals. Such an increase in the basal activity of VTA DA neurons [9,35] occurs through desensitization and up-regulation of nicotinic receptors and the long-term strengthening of glutamatergic synaptic transmission [51]. Here we show that acutely elevating VTA DA neuron activity using optogenetics is sufficient to mimic the behavioral alterations seen in mice under chronic nicotine exposure. Nicotine and optogenetic stimulation obviously act differently on DA neurons, yet they both lead to an increase in VTA DA neuron activity. Our optogenetic experiments confirmed once more that acutely modifying DA neuron activity did not change the animal's knowledge of reward probabilities (learning), but the way the animals used the values to form a decision strategy or policy. We thus link DA tonic neuromodulatory function and modifications of decision parameters (here β).

Variations in neuromodulatory functions, including those in the catecholamine and cholinergic systems, contribute to the process of individuation [52–54]. DA, and in particular from VTA DA neurons, has been linked to a cluster of traits (extraversion, novelty-seeking, etc.) conceptually related to reward-seeking [55,56]. However, despite the substantial attention paid to DA in personality neuroscience, and despite a clear association between modulations in dopaminergic function and variations in individual traits, defining which specific traits are influenced by DA remains a challenging task. Our data suggest that modification in basal VTA DA neuron activity can directly modify the expression of a specific trait: the exploit/explore tradeoff here estimated through the β variable. This result is reminiscent of the observations made from male mice living continuously in a large environment, which display idiosyncratic behavioral strategies during a decision-making task, and for which the exploration/exploitation balance was correlated with the activity of their DA system [54].

Nicotine exposure alters decision-making processes [6]. Non-contingency studies have previously shown that yoked nicotine exposure increases the incentive salience of non-nicotine stimuli [57], similar to the sensitization to ICSS rewards [47]. These studies suggest an essential role of contextual cues in smoking and the nicotine-induced increase in reward sensitivity. Neuroeconomics studies have also linked smoking with increased impulsivity (delay discounting task [11]), lack of counterfactual learning signals [58], and decreased behavioral flexibility (exploration in a dynamic bandit task [13]). Our results further reveal that nicotine exposure decreases exploration. In addition, we provide a mechanistic understanding of how reward processing may be altered at the level of the VTA in response to chronic nicotine. This insight is translationally valuable as nicotine-induced alterations in explore/exploit processes likely also have implications for the everyday life of smokers, particularly as they can increase vulnerability for addiction to other drugs of abuse and for behavioral disorders such as pathological gambling that rely on value-based decisions [7,59] and present a high comorbidity with tobacco addiction [60]. Our data underscore altered choice behaviors in smokers that likely participate in, but are not limited to, addiction [6]. Finally, such an

explore-exploit paradigm and archetypal analysis could be very useful to study the effects of other drugs of abuse on decision-making. Indeed, humans with methamphetamine [61] or alcohol use disorders [62] have been suggested to present alterations in bandit tasks, but human studies cannot disambiguate whether altered decision-making facilitates, or results from, drug use. Hence, preclinical studies are needed to dissect the causal mechanisms underlying alterations in economic decisions, and to understand the dynamics of drug users' profiles in general, and of smokers (or vape users) in particular.

Methods

Animals: Experiments were performed on adult C57Bl/6Rj DAT^{iCRE} and Wild-Type (Janvier Labs, France) mice. Male mice, from 8 to 16 weeks old, weighing 25-35 grams, were used for all the experiments. They were kept in an animal facility where temperature (20 ± 2°C) and humidity were automatically monitored and a circadian light cycle of 12/12-h light-dark cycle was maintained. All experiments were performed in accordance with the recommendations for animal experiments issued by the European Commission directives 219/1990, 220/1990 and 2010/63, and approved by Sorbonne University.

AAV production: AAV vectors were produced as previously described [63] using the cotransfection method and purified by iodixanol gradient ultracentrifugation [64]. AAV vector stocks were tittered by quantitative PCR (qPCR) [65] using SYBR Green (Thermo Fischer Scientific).

Intracranial self-stimulation electrode implantation: Mice were anaesthetized with a gas mixture of oxygen (1 L/min) and 1-3 % of isoflurane (Piramal Healthcare, UK), then placed into a stereotaxic frame (Kopf Instruments, CA, USA). After the administration of a local anesthetic (Lurocain, 0.1 mL at 0.67 mg/kg), a median incision revealed the skull which was drilled at the level of the Median Forebrain Bundle (MFB). A bipolar stimulating electrode (PlasticOne 2-channels, stainless steel, 10mm) for ICSS was then implanted unilaterally (randomized) in the brain (stereotaxic coordinates from bregma according to mouse after Paxinos atlas: AP -1.4 mm, ML ±1.2 mm, DV -4.8 mm from the brain). Dental cement (SuperBond, Sun Medical) was used to fix the implant to the skull. After stitching and administration of a dermal antiseptic, mice were then placed back in their home-cage and had, at least, 5 days to recover from surgery. An analgesic, buprenorphine solution at 0,015 mg/L (0,1 mL / 10 g), was delivered after the surgery and if necessary, the following recovering days. The efficacy of electrical stimulation was verified through the rate of acquisition during the deterministic setting (see behavioral methods).

Implantation of osmotic mini pumps: After 5 days of training in the deterministic setting (see behavioral methods), animals were anesthetized with a gas mixture of oxygen (1L/min) and 1-3 % of isoflurane (IsoVet, Piramal Healthcare, UK). After the administration of a local anesthetic, an incision was performed at the level of the interscapular zone, to subcutaneously implant an osmotic minipump (Model 2004, ALZET, CA, USA) containing 200 µL of either a solution of nicotine hydrogen tartrate salt (Sigma-Aldrich, USA) at a dose of 10 mg/kg/day (free base) or saline solution (H_2O with 0.9 % NaCl) for the control group. Both solutions were prepared in the laboratory. Minipumps delivered their content with a flow rate of 0.25 µL/hour over 28 days. The surgical wound was closed with surgical stitches. Animals had two days of rest to recover from the minipump surgery before going further with their behavioral training.

Virus injections and optogenetics experiments: DAT^{iCRE} mice were anaesthetized (Isoflurane 1-3%) and implanted with an ICSS electrode as described above. They were then injected unilaterally (randomized left/right

side and ipsi/contralateral side regarding the ICSS electrode) in the VTA (1 µL, coordinates from bregma: AP -3.1 mm; ML ±0.5 mm; DV -4.55 mm from the skull) with an adeno-associated virus (AAV5.Ef1α.DIO.hCatCh.YFP $2.46e^{12}$ - $6.53e^{13}$ vg/ml, AAV5.Ef1α.DIO.Jaws.eGFP $1.16e^{13}$ vg/ml or AAV5.Ef1α.DIO.YFP $6.89e^{13}$ or $9.10e^{13}$ vg/ml). A double-floxed inverse open reading frame (DIO) allowed to restrain the expression of CatCh (Ca^{2+}-translocating channelrhodopsin) or Jaws (red-shifted cruxhalorhodopsin) to VTA dopaminergic neurons of DAT^{iCRE} mice.

For optogenetic experiments on freely moving mice, an optical fiber (200 µm core, NA = 0.39, Thor Labs) coupled to a ferule (1.25 mm) was implanted just above the VTA ipsilateral to the viral injection (coordinates from bregma: AP -3.1 mm, ML ±0.5 mm, DV 4.4 mm), and fixed to the skull with dental cement (SuperBond, Sun Medical). The behavioral task began at least 4 weeks after virus injection to allow the transgene to be expressed in the target dopamine cells. An ultra-high-power LED (470 nm or 520 nm, Prizmatix) coupled to a patch cord (500 µm core, NA = 0.5, Prizmatix) was used for optical stimulation (output intensity of 10 mW). Optical stimulation was delivered continuously, starting 5 min before and continuing throughout the 5 min of ON sessions of the task. Excitatory opsin (CatCh) was stimulated using 470 nm light pulses of 5ms duration and 8 Hz frequency. Inhibitory opsin (Jaws) was stimulated using 520 nm light pulses of 500 ms duration and 0.5 Hz frequency. The experiment followed a schedule of paired ON and OFF days after the end of training phase (deterministic + probabilistic settings). The optical stimulation patch cord was plugged onto the ferrule during all experimental sessions (ON and OFF days) to habituate animals and control for latent experimental effects.

***Ex vivo* patch-clamp recordings of VTA DA neurons:** To verify the functional expression of the excitatory opsin CatCh and the inhibitory opsin Jaws, 10-12 week-old male DAT^{iCRE} mice were injected with the viruses described above. After 4 weeks, mice were deeply anesthetized with an intraperitoneal (IP) injection of a mix of ketamine/xylazine. Coronal midbrain sections (250 µm) were sliced using a Compresstome (VF-200; Precisionary Instruments) after intracardial perfusion of cold (4°C) sucrose-based artificial cerebrospinal fluid (SB-aCSF) containing (in mM): 125 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 5.9 $MgCl_2$, 26 $NaHCO_3$, 25 Sucrose, 2.5 Glucose, 1 Kynurenate (pH 7.2, 325 mOsm). After 10-60 min at 35°C for recovery, slices were transferred into oxygenated aCSF containing (in mM): 125 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 2 $CaCl_2$, 1 $MgCl_2$, 26 $NaHCO_3$, 15 Sucrose, 10 Glucose (pH 7.2, 325 mOsm) at room temperature for the rest of the day and individually transferred to a recording chamber continuously perfused at 2 ml/min with oxygenated aCSF. Patch pipettes (4–8 MΩ) were pulled from thin wall borosilicate glass (G150TF-3, Warner Instruments) using a micropipette puller (P-87, Sutter Instruments, Novato, CA) and filled with a KGlu based intra-pipette solution containing (in mM): 116 K-gluconate, 10-20 HEPES, 0.5 EGTA, 6 KCl, 2 NaCl, 4 ATP, 0.3 GTP and 2 mg/mL biocytin (pH adjusted to 7.2). Transfected VTA DA cells were visualized using an upright microscope coupled with a Dodt contrast lens and illuminated with a white light source (Scientifica). To characterize CatCh expression, a 460 nm LED (CoolLED) was used both for visualizing YFP positive cells (using a bandpass filter cube) and for optical stimulation through the microscope (1s continuous for light-evoked current in voltage-clamp mode and 8 Hz with 5 ms/pulses to drive neuronal firing in current-clamp mode). Regarding Jaws expression, continuous photostimulation (20 s), with a 525 nm, pE-2, CoolLED, was used

in current-clamp (-60 mV). Whole-cell recordings were performed using a patch-clamp amplifier (Axoclamp 200B, Molecular Devices) connected to a Digidata (1550 LowNoise acquisition system, Molecular Devices). Signals were low pass filtered (Bessel, 2 kHz) and collected at 10 kHz using the data acquisition software pClamp 10.5 (Molecular Devices). All the electrophysiological recordings were extracted using Clampfit (Molecular Devices) and analyzed with R.

***In vivo* juxtacellular recordings of VTA DA neurons:** Mice were deeply anaesthetized with chloral hydrate (8%), 400 mg/kg IP, supplemented as required to maintain optimal anesthesia throughout the experiment. The scalp was opened and a hole was drilled in the skull above the location of the VTA. Extracellular recording electrodes were constructed from 1.5 mm O.D. / 1.17 mm I.D. borosilicate glass tubings (Harvard Apparatus) using a vertical electrode puller (Narishige). Under microscopic control, the tip was broken to obtain a diameter of approximately 1 µm. The electrodes were filled with a 0.5% NaCl solution containing 1.5% of Neurobiotin tracer (AbCys) yielding impedances of 6-9 MΩ. Electrical signals were amplified by a high-impedance amplifier (Axon Instruments) and monitored through an audio monitor (A.M. Systems Inc.). The signal was digitized, sampled at 25 kHz and recorded using Spike2 software (Cambridge Electronic Design) for later analysis. The electrophysiological activity was sampled in the central region of the VTA (coordinates: between 3.1 to 4 mm posterior to bregma, 0.3 to 0.7 mm lateral to midline, and 4 to 4.8 mm below brain surface). Individual electrode tracks were separated from one another by at least 0.1 mm in the horizontal plane. Spontaneously active DA neurons were identified based on previously established electrophysiological criteria [66,67]

Fluorescence immunohistochemistry: After euthanasia, induced by IP injection of euthasol (0.1 mL per 30 g at 150 mg/kg) or by paraformaldehyde (PFA) intra-cardiac perfusion, brains were rapidly removed and fixed in 4% PFA. Following a period of fixation at 4°C, serial 60-µm sections were cut from the midbrain with a vibratome. Immunohistochemistry was performed as follows: free-floating VTA brain sections were incubated 1h at 4°C in a blocking solution of phosphate-buffered saline (PBS) containing 3% Bovine Serum Albumin (BSA, Sigma A4503) and 0.2% Triton X-100 and then incubated overnight at 4°C with a mouse anti-tyrosine hydroxylase antibody (TH, Sigma, T1299) at 1:200 dilution in PBS containing 1.5% BSA and 0.2% Triton X-100. The following day, sections were rinsed with PBS and then incubated for 3h at 22–25 °C with Cy3-conjugated anti-mouse (Jackson ImmunoResearch, 715-165-150) at 1:200 dilution in a solution of 1.5% BSA in PBS, respectively. After three rinses in PBS, slices were wet-mounted using Prolong Gold Antifade Reagent (Invitrogen, P36930). Microscopy was carried out with a fluorescent microscope Leica DMR, and images captured in grey level using MetaView software (Universal Imaging Corporation) and colored post-acquisition with ImageJ.

For the optogenetic experiments on DAT^{iCRE} mice, an immunohistochemical identification of the transfected neurons was performed as described above, with an addition of chicken anti-eYFP antibodies (Life technologies Molecular Probes, A-6455), at 1:500 dilution. A goat-anti-chicken AlexaFluor 488 secondary antibody (711-225-152, Jackson ImmunoResearch) at 1:500 dilution (Life Technologies) was then used in a solution of 1.5% BSA in

PBS. Neurons co-labelled for TH and YFP in the VTA allowed to confirm their neurochemical phenotype and the transfection success.

Intracranial self-stimulation (ICSS) bandit task

Behavioral set up: The ICSS bandit task took place in a circular open field with a diameter of 67 cm. Three explicit square-shaped marks (1x1 cm) were placed in the open field, forming an equilateral triangle (side = 35 cm). Entry in the circular zones (diameter = 6 cm) around each mark was associated with the delivery of a rewarding ICSS stimulation. Experiments were performed using a video camera, connected to a video-tracking system, out of sight of the experimenter. A LabVIEW (National Instruments) application precisely tracked and recorded the animal's position with a camera (20 frames/s). When a mouse was detected in one of the circular rewarding zones, an electrical stimulator received a TTL signal from the software application and generated a 200 ms-train of 0.5-ms biphasic square waves pulsed at 100 Hz (20 pulses per train). ICSS intensity was adjusted, within a range of 20 to 200 µA, during training (see training settings) and then kept constant, so that mice would achieve between 50 and 150 visits per session (5min duration) for two successive sessions, and then kept constant for all the experiment. Mice with insufficient scores in the deterministic and probabilistic setting (<40 visits despite increasing the maximum intensity of 200 µA) were excluded.

Training setting: The training consisted of two settings: the deterministic setting (DS) and the probabilistic setting (PS), both consisting of at least 10 daily sessions of 5 min. In the deterministic setting, all zones were associated with an ICSS delivery (P = 100%). However, two consecutive rewards could not be delivered on the same target, which motivates mice to alternate between targets. In the probabilistic setting, the zones were associated with three different probabilities (P = 25%, P = 50%, P = 100%) to obtain an ICSS stimulation. The probabilities locations were pseudo-randomly assigned per mouse.

Data acquisition per experimental group: Different experimental groups underwent the ICSS bandit task. Firstly, locomotion and choice behavior of the mice, which had been implanted with osmotic mini-pumps (Saline = 23, Nicotine = 27), were analyzed and compared between the last two days of both training settings. For optogenetics experiments, the DAT^{iCRE} mice (n = 21) completed the training, followed by a schedule of paired sessions with photo-stimulation (ON) alternated with days without photostimulation (OFF). The control animals (n = 55) were obtained by pooling together mice implanted with a saline mini-pump (n = 23) and non-implanted mice (n = 32). Figure 1 used data from the non-implanted mice group. Figure 2,3,4 used the pooled control group.

Behavioral measures: For all of those groups, the following measures were analyzed and compared in the probabilistic setting, as well as in the deterministic setting for the saline vs nicotine experiment: i) number of visits, ii) time-to-goal, iii) choice repartition (proportion of visits p_{25}, p_{50} and p_{100}), iv) percentage of directional changes (n^{th} visit = n^{th} visit+2). Furthermore, the ICSS bandit task can be seen as a Markovian decision process. Every transition between zones can be considered as a binary choice between two probabilities, since the occupied zone cannot

be reinforced twice in a row. The sequence of choices per session is summarized by the proportional result of the sum of three specific binary choices (or gambles, i.e., total visits zone 1/total visits zone 1+2). The three gambles (G) were named after the point on which the mouse is positioned at the time of the choice: G_{25} = 100 % vs 50 %, G_{100} = 50 % vs 25 % and G_{50} = 100 % vs 25 %. The target choice in these gambles reflects the balance between exploitative (choosing the most valuable option) and exploratory (choosing the least valuable option) choices. With a softmax based decision-making model fitted in the laboratory, we computed three parameters: the value sensitivity or inverse temperature (the power to discriminate between values in a binary choice), the uncertainty bonus (the preference for expected uncertainty, considering the reward variance of every option in a binary choice) and the motor cost to do a directional change (a decrease in the target value if it requires to go back to the previous target).

Modeling: Decision-making models determined the probability P_i of choosing the next state i, as a function (the "choice rule") of a "decision variable". Because mice could not return to the same rewarding target, they had to choose between the two remaining ones. Accordingly, we modeled decisions between two alternatives labelled A and B and used a softmax choice rule defined by $P_A = 1 / (1+e^{-ß(vA-vB)})$ where β is an inverse temperature parameter reflecting the sensitivity of choice to the difference between decision variables and V_i the value of an option. The value *V* of an option is modelled as the expected (average) reward + expected uncertainty + U-turn cost [19,33]. This compound value is then nested in the softmax choice rule, given a 6*3 matrix that described the probability of a choice between A, B and C (the three targets) depending on the two previous choices. As an example, in the probability to choose (A, B, C) after performing the sequence BA, the value is given by $(0, p_b + \varphi * p_b*(1-p_b) - \kappa, p_c + \varphi * p_c*(1-p_c))$ while after the sequence CA the value is given by $(0, p_b+\varphi * p_b*(1-p_b) , p_c+\varphi * p_c*(1-p_c) - \kappa)$ (same for AB, CB and AC, BC). The free parameters of the model (β, φ, κ) were fitted by maximizing the data likelihood. Given a sequence of choice $c = c_{1..T}$, data likelihood is the product of their probability (given by Equation 1) [68]. We used the *optim* function in R to perform the fits, with the constraints that $\beta \in]0,10]$, $\varphi \in]0,5]$ and $\kappa \in]0,5]$.

In the model, the β parameter reflects how much the difference in total value between the two options (ΔV) translates into more or less preference for the best option in a given gamble. With a small β, choices are little sensitive to ΔV. With a null β, both options have the same probability to be selected in each gamble (G100 = G50 = G25 = 50%), leading to equal global distribution of visits (p100 = p50 = p25 = 33%), independently of their respective value. On the contrary, a large β indicates a high sensitivity to ΔV, with an infinite beta indicating that options associated with higher reward probabilities are always selected (G100 = G50 = 100% and G25 would not even exist considering that animals would never visit p25%; p100 = p50 = 50% and p25 = 0%).

Model comparison: To compare models [68], we used the Bayesian Information Criterion (BIC) to correct the raw likelihoods for the number of free parameters fit. BIC scores were aggregated across mice (Figure 4b). M1 and M2 are nested cases of M3. In M3, $\beta>0$, $\varphi>0$ and $\kappa>0$. In M1, β and $\varphi=0$, so that choices are only driven by a motor cost $\kappa>0$. In M2, $\varphi=0$, corresponding to animals that do not take uncertainty into account. A likelihood ratio test

was used to estimate the probability of the observed data under the null hypothesis that these data are generated by the simplest model. For that we compute d, twice the difference in log likelihoods of M2 or M1 with M3. The probability of a difference d follows a chi-square distribution with a number of degrees of freedom n equal to the difference of parameters number between M3 and M1 or M2 (i.e n=1 or 2) [68].

Statistical analysis: All statistical analyses were computed using R (The R Project, version 4.0.0) and Python with custom programs. Results were plotted as a mean ± s.e.m. The total number (n) of observations in each group and the statistics used are indicated in figure legends. Classical comparisons between means were performed using parametric tests (Student's T-test, or ANOVA for comparing more than two groups when parameters followed a normal distribution (Shapiro test $P > 0.05$), and non-parametric tests (here, Wilcoxon or Mann-Whitney) when the distribution was skewed. Multiple comparisons were corrected using a sequentially rejective multiple test procedure (Holm). Probability distributions were compared using the Kolmogorov–Smirnov (KS) test, and proportions were evaluated using a chi-squared test (χ^2). All statistical tests were two-sided except for the optogenetic experiment (Figure 5) where statistical tests were paired and one-sided (we test hypotheses driven by nicotine effect and model). $P > 0.05$ was considered not to be statistically significant. For archetypal analysis, all computations and graphics have been done using the statistical software R and the archetype package (version 2.2-0.1). Briefly, given an n × m matrix representing a multivariate data set with n observations (n = number of animals) and m attributes (here $m = 7$, consisting of the directional changes, the target distributions (3 values) and the three gambles (see data Figure 1 c-e)), the archetypal analysis finds the matrix Z of k m-dimensional archetypes (k is the number of archetypes). Z is obtained by minimizing $|| X-\alpha Z^T ||_2$, with α the coefficients of the archetypes ($\alpha_{i,1..k} \geq 0$ and $\sum \alpha_{i,1..k} = 1$), and $||.||_2$ a matrix norm. The archetype is also a convex combination of the data points $Z=X^T\delta$, with $\delta \geq 0$ and their sum must be 1 [69]. The α-coefficient depicts the relative archetypal composition of a given observation. For $k = 3$ archetypes and an observation i, $\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3} \geq 0$ and $\alpha_{i,1} + \alpha_{i,2} + \alpha_{i,3} = 1$. A ternary plot can then be used to visualize data. ($\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,2}$) are used to assign individual behavior to its nearest archetype (i.e, k max($\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}$)). $\alpha_{i,j}$ are also used as variable to estimate population archetypal composition. For figure 4e, archetypal composition ($0 \leq \alpha_{i,j} \leq 1$) was binned into five intervals. Pure archetype corresponds to 1, the archetypal composition decreases linearly with increasing distance from the archetype, 0 correspond to points on the opposite side.

Data Availability: All the data and code are available from the corresponding authors upon reasonable request.

Code Availability: Results were generated using code written in R and Python. All codes used to run the analysis are available from the authors upon request.

References

1. Marti, F. et al. Smoke extracts and nicotine, but not tobacco extracts, potentiate firing and burst activity of ventral tegmental area dopaminergic neurons in mice. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 36, 2244–2257 (2011).

2. Fowler, C. D., Arends, M. A. & Kenny, P. J. Subtypes of nicotinic acetylcholine receptors in nicotine reward, dependence, and withdrawal: evidence from genetically modified mice. Behavioural pharmacology 19, 461–484 (2008).

3. Stolerman, I. P. & Jarvis, M. J. The scientific case that nicotine is addictive. Psychopharmacology (Berl) 117, 2-10-discussion 14-20 (1995).

4. Lüscher, C. & Malenka, R. C. Drug-evoked synaptic plasticity in addiction: from molecular changes to circuit remodeling. Neuron 69, 650–663 (2011).

5. Association, A. P. Diagnostic and statistical manual of mental disorders (DSM-5®). (2013).

6. Naudé, J., Dongelmans, M. & Faure, P. Nicotinic alteration of decision-making. Neuropharmacology 96, 244–254 (2015).

7. Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L. & Platt, M. L. A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 42, 1931–1939 (2017).

8. Picciotto, M. R., Lewis, A. S., Schalkwyk, G. I. van & Mineur, Y. S. Mood and anxiety regulation by nicotinic acetylcholine receptors: A potential pathway to modulate aggression and related behavioral states. Neuropharmacology 96, 235–243 (2015).

9. Morel, C. et al. Nicotinic receptors mediate stress-nicotine detrimental interplay via dopamine cells' activity. Molecular Psychiatry 23, 1597–1605 (2017).

10. Nguyen, C. et al. Nicotine inhibits the VTA-to-amygdala dopamine pathway to promote anxiety. Neuron (2021) doi:10.1016/j.neuron.2021.06.013.

11. Locey, M. L. & Dallery, J. Isolating behavioral mechanisms of intertemporal choice: nicotine effects on delay discounting and amount sensitivity. Journal of the experimental analysis of behavior 91, 213–223 (2009).

12. Viñals, X. et al. Overexpression of α3/α5/β4 nicotinic receptor subunits modifies impulsive-like behavior. Drug and Alcohol Dependence 122, 247–252 (2012).

13. Addicott, M. A., Pearson, J. M., Froeliger, B., Platt, M. L. & McClernon, F. J. Smoking automaticity and tolerance moderate brain activation during explore-exploit behavior. Psychiatry Research 224, 254–261 (2014).

14. Addicott, M. A., Pearson, J. M., Wilson, J., Platt, M. L. & McClernon, F. J. Smoking and the bandit: A preliminary study of smoker and nonsmoker differences in exploratory behavior measured with a multiarmed bandit task. Experimental and Clinical Psychopharmacology 21, 66–73 (2013).

15. Levine, A. et al. Molecular mechanism for a gateway drug: epigenetic changes initiated by nicotine prime gene expression by cocaine. Science translational medicine 3, 107ra109 (2011).

16. Wyart, V., Sciences, E. K. C. O. in B. & 2016. Choice variability and suboptimality in uncertain environments. Personality and Individual Differences 11, 109–115 (2016).

17. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical transactions of the Royal Society of London Series B, Biological sciences 362, 933–942 (2007).

18. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. Journal of experimental psychology. General 143, 2074–2081 (2014).

19. Belkaid, M. et al. Mice adaptively generate choice variability in a deterministic task. Communications Biology 3, 1–9 (2020).

20. Berlyne, D. E. Curiosity and exploration. Science (New York, N.Y.) 153, 25–33 (1966).

21. Redish, A. D., Jensen, S. & Johnson, A. A unified framework for addiction: Vulnerabilities in the decision process. The Behavioral and brain sciences 31, 415–37; discussion 437-87 (2008).

22. Schultz, W. Multiple dopamine functions at different time courses. Annual review of neuroscience 30, 259–288 (2007).

23. Lüscher, C., Robbins, T. W. & Everitt, B. J. The transition to compulsion in addiction. Nature Reviews Neuroscience 21, 1–17 (2020).

24. Kalivas, P. W. & Volkow, N. D. The Neural Basis of Addiction: A Pathology of Motivation and Choice. Am J Psychiat 162, 1403–1413 (2005).

25. Mizumori, S. J. Y. & Jo, Y. S. Homeostatic regulation of memory systems and adaptive decisions. Hippocampus 23, 1103–1124 (2013).

26. Cagniard, B. et al. Dopamine scales performance in the absence of new learning. Neuron 51, 541–547 (2006).

27. Westbrook, A. & Braver, T. S. Dopamine Does Double Duty in Motivating Cognitive Effort. Neuron 89, 695–710 (2016).

28. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl) 191, 507–520 (2007).

29. Beeler, J. A., Daw, N., Frazier, C. R. M. & Zhuang, X. Tonic dopamine modulates exploitation of reward learning. Frontiers in Behavioral Neuroscience 4, 170 (2010).

30. Frank, M. J., Doll, B. B., Oas-Terpstra, J. & Moreno, F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience 12, 1062–1068 (2009).

31. Humphries, M. D., Khamassi, M. & Gurney, K. Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. Frontiers in neuroscience 6, 9 (2012).

32. Cinotti, F. et al. Dopamine blockade impairs the exploration-exploitation trade-off in rats. Scientific Reports 9, 6770 (2019).

33. Naudé, J. et al. Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nature Neuroscience 19, 471–478 (2016).

34. Epping-Jordan, M. P., Watkins, S. S., Koob, G. F. & Markou, A. Dramatic decreases in brain reward function during nicotine withdrawal. Nature 393, 76–79 (1998).

35. Tolu, S. et al. Nicotine enhances alcohol intake and dopaminergic responses through β2* and β4* nicotinic acetylcholine receptors. Scientific Reports 7, 45116 (2017).

36. Cutler, A. & Breiman, L. Archetypal Analysis. Technometrics 36, 338–347 (1994).

37. Hart, Y. et al. Inferring biological tasks using Pareto analysis of high-dimensional data. Nature Methods 12, 233-5–3 p following 235 (2015).

38. Besson, M. et al. Long-term effects of chronic nicotine exposure on brain nicotinic receptors. Proceedings of the National Academy of Sciences of the United States of America 104, 8155–8160 (2007).

39. Kleinlogel, S. et al. Ultra light-sensitive and fast neuronal activation with the Ca^{2}+-permeable channelrhodopsin CatCh. 14, 513–518 (2011).

40. Chuong, A. S. et al. Noninvasive optical inhibition with a red-shifted microbial rhodopsin. 1123–1129 (2014).

41. Faure, P., Tolu, S., Valverde, S. & Naudé, J. Role of nicotinic acetylcholine receptors in regulating dopamine neuron activity. Neuroscience 282C, 86–100 (2014).

42. Heilbronner, S. R. & Hayden, B. Y. Contextual factors explain risk-seeking preferences in rhesus monkeys. Frontiers in neuroscience 7, 7 (2013).

750 43. Musall, S., Urai, A. E., Sussillo, D. & Churchland, A. K. Harnessing behavioral diversity to understand neural
751 computations for cognition. Curr Opin Neurobiol 58, 229–238 (2019).

752 44. Calipari, E. S. et al. Dopaminergic dynamics underlying sex-specific cocaine reward. Nature Communications
753 8, 13877–15 (2017).

754 45. Rezvani, A. H. & Levin, E. D. Cognitive effects of nicotine. Biol Psychiat 49, 258–267 (2001).

755 46. Hernandez, G., Trujillo-Pisanty, I., Cossette, M.-P., Conover, K. & Shizgal, P. Role of dopamine tone in the
756 pursuit of brain stimulation reward. Journal of Neuroscience 32, 11032–11041 (2012).

757 47. Kenny, P. J. & Markou, A. Nicotine self-administration acutely activates brain reward systems and induces a
758 long-lasting increase in reward sensitivity. Neuropsychopharmacology : official publication of the American
759 College of Neuropsychopharmacology 31, 1203–1211 (2006).

760 48. Grace, A. A. & Bunney, B. S. The control of firing pattern in nigral dopamine neurons: burst firing. The Journal
761 of neuroscience : the official journal of the Society for Neuroscience 4, 2877–2890 (1984).

762 49. Tritsch, N. X. & Sabatini, B. L. Dopaminergic Modulation of Synaptic Transmission in Cortex and Striatum.
763 Neuron 76, 33–50 (2012).

764 50. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. Nature
765 Neuroscience 1–10 (2013) doi:10.1038/nn.3413.

766 51. Juarez, B. et al. Midbrain circuit regulation of individual alcohol drinking behaviors in mice. Nature
767 Communications 8, 2220 (2017).

768 52. Stern, S., Kirst, C. & Bargmann, C. I. Neuromodulatory Control of Long-Term Behavioral Patterns and
769 Individuality across Development. Cell 171, 1–25 (2017).

770 53. MacDonald, S. W. S., Nyberg, L. & Bäckman, L. Intra-individual variability in behavior: links to brain structure,
771 neurotransmission and neuronal activity. Trends in Neurosciences 29, 474–480 (2006).

772 54. Torquet, N. et al. Social interactions impact on the dopaminergic system and drive individuality. Nature
773 Communications 9, 3081 (2018).

774 55. Smillie, L. D. & Wacker, J. Dopaminergic foundations of personality and individual differences. Frontiers in
775 Human Neuroscience 8, 874 (2014).

776 56. DeYoung, C. G. The neuromodulator of exploration: A unifying theory of the role of dopamine in personality.
777 Frontiers in Human Neuroscience 7, 762 (2013).

778 57. Palmatier, M. I. et al. Dissociating the primary reinforcing and reinforcement-enhancing effects of nicotine
779 using a rat self-administration paradigm with concurrently available drug and environmental reinforcers.
780 Psychopharmacology (Berl) 184, 391–400 (2006).

781 58. Chiu, P. H., Lohrenz, T. M. & Montague, P. R. Smokers' brains compute, but ignore, a fictive error signal in a
782 sequential investment task. Nature Neuroscience 11, 514–520 (2008).

783 59. Addicott, M. A., Pearson, J. M., Kaiser, N., Platt, M. L. & McClernon, F. J. Suboptimal foraging behavior: a
784 new perspective on gambling. Behavioral neuroscience 129, 656–665 (2015).

785 60. McGrath, D. S. & Barrett, S. P. The comorbidity of tobacco smoking and gambling: a review of the literature.
786 Drug and alcohol review 28, 676–681 (2009).

787 61. Harlé, K. M. et al. Altered Statistical Learning and Decision-Making in Methamphetamine Dependence:
788 Evidence from a Two-Armed Bandit Task. Frontiers in psychology 6, 1910 (2015).

789 62. Morris, L. S. et al. Biases in the Explore-Exploit Tradeoff in Addictions: The Role of Avoidance of Uncertainty.
790 Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 41, 940–
791 948 (2016).

792 63. Khabou, H. et al. Noninvasive gene delivery to foveal cones for vision restoration. JCI insight 3, D358 (2018).

793 64. Choi, V. W., Asokan, A., Haberman, R. A. & Samulski, R. J. Production of recombinant adeno-associated viral
794 vectors. Current protocols in human genetics Chapter 12, Unit 12.9-12.9.21 (2007).

795 65. Aurnhammer, C. et al. Universal real-time PCR for the detection and quantification of adeno-associated virus
796 serotype 2-derived inverted terminal repeat sequences. Human gene therapy methods 23, 18–28 (2012).

797 66. Exley, R. et al. Distinct contributions of nicotinic acetylcholine receptor subunit alpha4 and subunit alpha6 to
798 the reinforcing effects of nicotine. Proceedings of the National Academy of Sciences of the United States of
799 America 108, 7577–7582 (2011).

800 67. Ungless, M. A. & Grace, A. A. Are you or aren't you? Challenges associated with physiologically identifying
801 dopamine neurons. Trends in Neurosciences 35, 422–430 (2012).

802 68. Daw, N. D. Trial-by-trial data analysis using computational models. in Decision Making, Affect, and Learning
803 3–38 (Oxford University Press, 2011).

804 69. Eugster, M. J. A. & Leisch, F. From Spider-Man to Hero - Archetypal Analysis in R. Journal of Statistical
805 Software 30, 1–23 (2009).
806

Acknowledgements:

We are grateful to the animal facilities (IBPS), Camille Robert and Paris Vision Institute AAV production facility for viral production and purification. This work was supported by the Centre National de la Recherche Scientifique CNRS UMR 8246, INSERM U1130, the Foundation for Medical Research (FRM, Equipe FRM DEQ2013326488 to P.F), FRM FDT201904008060 (to SM), the French National Cancer Institute Grant TABAC-16-022 et TABAC-19-020 (to P.F.), French state funds managed by the ANR (ANR-16 Nicostress, ANR -17 SNP-Nic, ANR-20 Nicado to PF, ANR-19 Vampire to FM) and The LabEx Bio-Psy (to P.F). MLD, RDC and SM were the recipients of a fourth-year PhD fellowship from FRM (FDT20160435171, FDT20170437427 and FDT201904008060), CN was recipient of a doctoral fellowship from the Labex Bio-Psy, DL was recipient of a post-doctoral Fellowship from the Labex Bio-Psy, and LMR was supported by a NIDA–Inserm Postdoctoral Drug Abuse Research Fellowship."

Author contributions:

MD, RDC, CN and MC contributed equally to this work. PF and MD designed the study. MD, RDC, CN, MC, TAY, EKD, RB, EB, BH, and NT performed the behavioral experiments. MD, RDC, SM and NT performed the minipumps implantations. JN, DL and SD contributed to setup developments. CN, SM, RDC, DL and FM performed electrophysiological recordings. MD, RDC, CN, MC, EKD, RB, TAY, EB, NT and JN performed the surgeries and virus injections. CN, SM, performed the immunohistochemistry experiments. DD provided the viruses. JN and PF developed the model. AM developed the optogenetic setup. MD, RDC, CN, MC, SM, JN, FM and PF analyzed the data. PF wrote the manuscript with inputs from MD, RDC, CN, MC, LMR, JN, FM and AM.

Competing Interests:

The authors declare no competing financial interests.

Figure 1: Mice exhibited suboptimal behavior and exploratory choices in a spatial version of a multi-armed bandit task with probabilistic settings

a Mice were implanted unilaterally with bipolar stimulation electrodes to deliver electrical stimulation at the level of the medial forebrain bundle in order to support intracranial self-stimulation (ICSS) behavior. *Right*: A coronal section of the mouse brain illustrating a representative electrode positioned in the MFB at Bregma -1.4 mm AP. **b** Schematic of the behavioral paradigm: mice are placed in a circular open-field (interior diameter = 80 cm), with three equidistant targets (A, B, and C - labelled on the open field floor) that are associated with a given probability ($P_{A,}$ $P_{B, or}$ P_C) of ICSS reward delivery when the animal is detected in a 60 mm zone around the target. **c** Sample trajectories for one mouse under the deterministic setting (DS) of the task, in which each of the three targets were rewarded by an ICSS with P = 100% (left panel, blue), and in the probabilistic setting (PS), in which the three targets were associated with distinct probabilities of ICSS delivery (P_A = 100, P_B = 50 and P_C = 25 %) (right panel, black). Two stimulations could not be delivered consecutively in the same zone, therefore animals learned to alternate between targets with a circular pattern in the deterministic setting (blue), and a less stereotyped pattern in the probabilistic setting (black). **d** Comparison of the percentage of directional changes during deterministic (DS, blue) and probabilistic setting (PS, black) (Wilcoxon signed rank test, $p < 0.001$, n = 33). **e** Repartition of visits to the three targets. Under the deterministic setting (blue), animals distributed uniformly their choices of visiting each of the three options (around 33%, Friedman rank sum test, p = 0.82). During the probabilistic setting (PS, black), animals reorganized their behavior and visited more frequently options with greater probabilities of reward (Friedman rank sum test, $p < 0.001$, and paired Wilcox Test $p < 0.001$ for the three comparisons, n = 33). **f** Probability to choose the option with the highest probability of reward for the three possible gambles: G_{100} = choice of 50% over 25%, G_{25} = choice of 100% over 50 and G_{50} = choice of 100% over 25%. Red asterisk: Comparison with a true mean of 0.5 (One Sample t-test with Holm correction, n = 33) for G_{100} (p = 0.026), G_{25} (p = 0.92) and G_{50} ($p < 0.001$). Black asterisk: Paired comparison (Paired t-test with Holm correction, n = 33) for G_{100}-G_{25} (p = 0.03), G_{100}-G_{50} (p = 0.048) and G_{25}-G_{50} (p = 0.025).

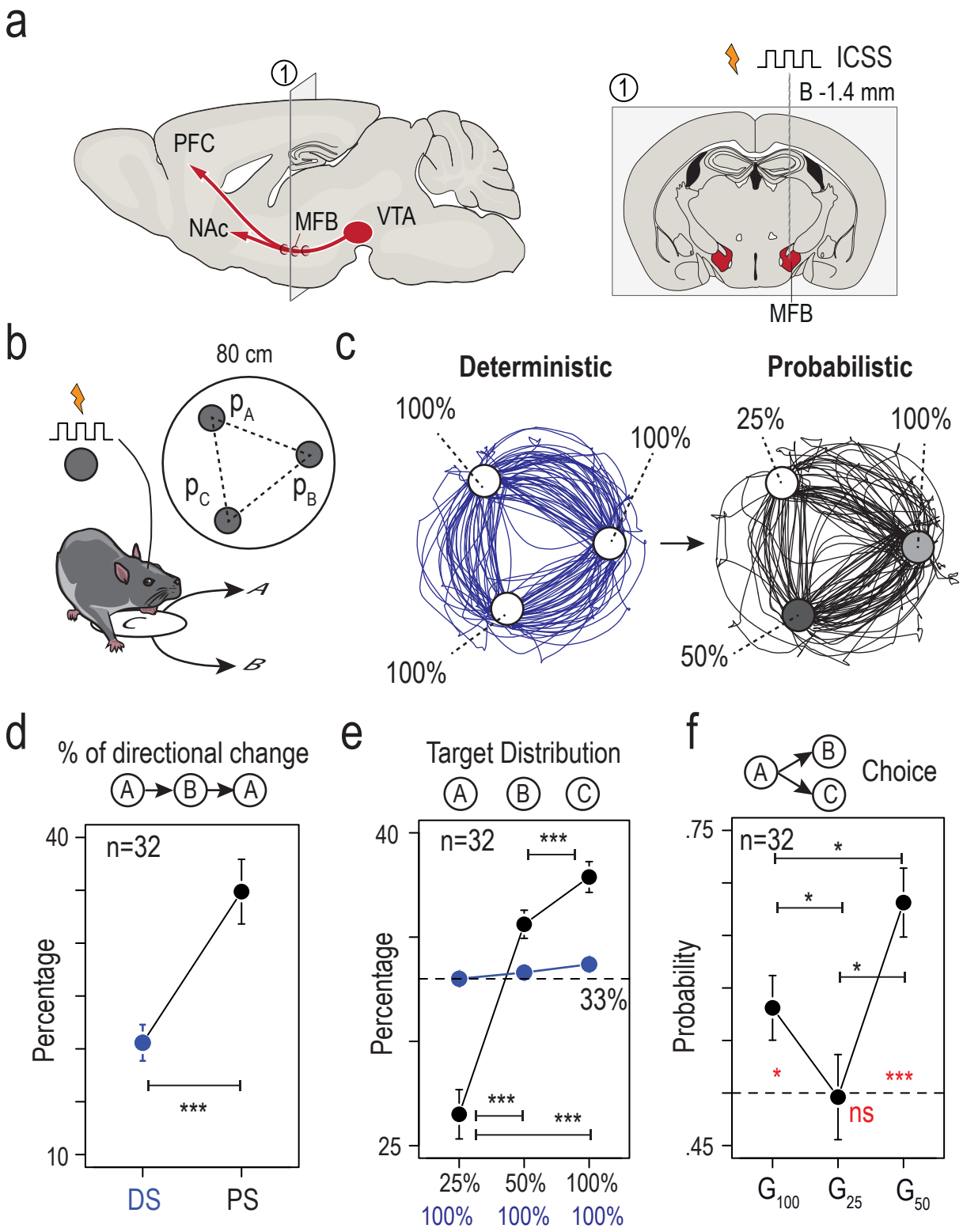

a
PFC
NAc
MFB
VTA
ICSS
B -1.4 mm
MFB
b
80 cm
p_A
p_B
p_C
A
B
C
c
Deterministic
Probabilistic
100%
100%
100%
25%
100%
50%
d
% of directional change
A → B → A
n=32
Percentage
40
10
***
DS
PS
e
Target Distribution
A B C
n=32
***
Percentage
40
25
33%
***
***
25% 50% 100%
100% 100% 100%
f
A B C
Choice
n=32
Probability
.75
.45
*
*
*
*
***
ns
G_{100} G_{25} G_{50}

Figure 1

Figure 2: Chronic exposure to nicotine altered both spontaneous DA activity and choice strategies.

a *Left*: Timeline of the task. Subcutaneous osmotic mini-pumps delivering nicotine (10 mg/kg/day), or saline for control animals, were implanted on day 5 of the deterministic setting (DS). *Right*: Sample trajectories at the end of the probabilistic setting (PS) for a mouse under chronic nicotine (Nic, in pink) and a mouse naive to nicotine (Ctl, in black). **b** *Left*: Representative in vivo electrophysiological recordings of VTA DA neurons after chronic saline (Sal, black) or nicotine (Nic, red) exposure. *Right*: The firing frequency and bursting activity of VTA DA neurons were compared between two sets of conditions: saline (n = 123) versus nicotine minipump (n = 62), and saline minipump + ICSS (n = 78) versus nicotine minipump + ICSS (n = 76) after completion of the probabilistic setting. All electrophysiological experiments were performed after 24 $\pm$ 2 days of Sal or Nic (10 mg/kg/day) exposure. Nicotine exposure increased both DA neuron firing frequency (two-way ANOVA, nicotine effect $F_{(1,335)} = 72.42$, $p < 0.001$) and bursting activity ($F_{(1,335)} = 25.39$, $p < 0.001$), with or without ICSS. This increase was observed between the Sal and Nic minipump-only conditions (*post hoc* Tukey HSD, firing frequency $p < 0.001$, bursting activity $p < 0.001$), as well as after Nic minipump + ICSS compared to Sal minipump + ICSS (*post hoc* Tukey HSD, firing frequency $p < 0.001$, bursting activity $p = 0.02$). Mean firing frequency was increased after ICSS in both the Sal and Nic groups (two-way ANOVA, ICSS effect $F_{(1,335)} = 21.53$, $p < 0.001$), but bursting activity was unchanged after ICSS ($F_{(1,335)} = 1.02$, $p= 0.31$). No interaction effect was observed for firing frequency ($F_{(1,335)} = 1.02$, $p = 0.31$) nor bursting activity ($F_{(1,335)} = 0.65$, $p = 0.42$). **c-f** Comparison between mice exposed to chronic nicotine (Nic, in pink, n = 27) and control mice (Ctl, n = 55) at the end of the deterministic setting, regarding **c** the target repartition (i.e., P_A, P_B and P_C, $p>0.05$), **d** the percentage of directional changes (Student's t-test, $p>0.05$), **e** the number of rewards (Student's t-test, ***$p < 0.001$) and **f** the cumulative distribution of the average time-to-goal (KS test, **$p < 0.01$). **g-j** Comparison between mice exposed to chronic nicotine (Nic, in pink, n = 27) and control mice (Ctl, n = 55) at the end of the probabilistic setting, regarding: **g** the target repartition. Nic mice visited more often the options with a higher reward probability (i.e. P_{50} and P_{100}) and less often the option with the lowest probability (P_{25}) in comparison to control mice (student t-test with Holm correction for multiple comparisons, **$p = 0.006$, *$p = 0.011$, *$p = 0.012$, respectively). **h** Percentage of directional changes (Student's t-test, *$p = 0.02$); **i** Probability of making the exploitative choice (i.e., the one with the highest probability of reward) for the three possible gambles for nicotine and control mice (Student's t-test with Holm correction for multiple comparisons, *$p = 0.03$) and **j** the cumulative distribution of the average time-to-goal (KS test, **$p < 0.01$).

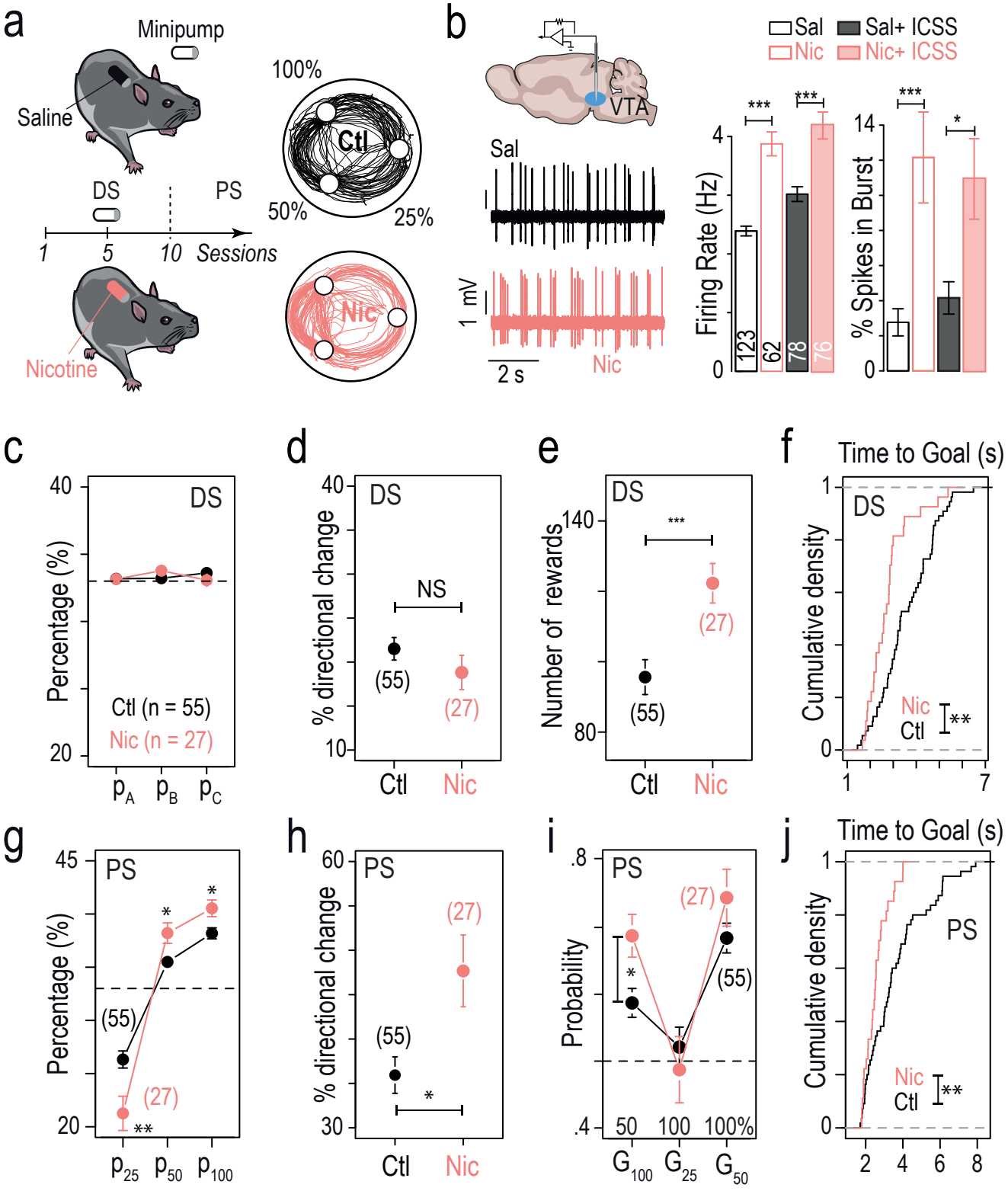

a
Minipump
Saline
DS
PS
1
5
10
Sessions
Nicotine
100%
Ctl
50%
25%
Nic
b
VTA
Sal
1 mV
2 s
Nic
Sal
Sal+ ICSS
Nic
Nic+ ICSS
Firing Rate (Hz)
% Spikes in Burst
123
62
78
76
c
DS
Percentage (%)
Ctl (n = 55)
Nic (n = 27)
d
DS
% directional change
NS
(55)
(27)
Ctl
Nic
e
DS
Number of rewards
(55)
(27)
Ctl
Nic
f
Time to Goal (s)
DS
Cumulative density
Nic
Ctl
g
PS
Percentage (%)
(55)
(27)
h
PS
% directional change
(55)
(27)
Ctl
Nic
i
PS
Probability
(27)
(55)
50
100
100%
j
Time to Goal (s)
PS
Cumulative density
Nic
Ctl

Figure 2

Figure 3: Mice exhibited inter-individual differences in choice strategies which were differentially affected by chronic nicotine exposure

a *Left*: Sample trajectories in the probabilistic setting, corresponding to different choice strategies, a circular strategy (*top*) and a gain-optimizing strategy (*bottom*). *Right*: A mouse using a purely circular strategy (top, descending Des or ascending Asc) in the probabilistic setting will tend to a 58.3 % success rate, whereas a mouse that always avoids p_{25} and alternates between p_{100} and p_{50} (bottom) will reach 75 % of success rate. **b** Correlation between the success rate and the percentage of directional changes. Mice displayed a strong inter-individual variability in their choice strategy but, overall, the higher the percentage of directional change, the higher the success rate (regression line in black). The red line indicates the linear correlation passing through two theoretical points: {0% directional changes; 58.3 % success rate} and {100 % directional changes; 75 % success rate}. **c** Correlation between the percentage of directional changes for two consecutive sessions. This measure showed a strong stability between consecutive sessions, indicating that the decision strategy was conserved across time for a given individual. **d-e** Archetypal analysis of the choice strategies based on 7-dimensional data space: i) the % of directional changes, ii) the gambles G_{100}, G_{25} and G_{50}, and iii) the distribution of choices between p_{25}, p_{50}, and p_{100}. Analysis was performed on n = 82 mice (pooled control and nicotine mice). **d** Plot of the three archetypal solutions, gain-maximizers (GO), descending (Des) and ascending (Asc), and their 7 basic variables used in this analysis. **e** *Left*: Visualization of the α coefficients using a ternary plot. Each point represents the projection of an individual onto the plane defined by a triangle where the three apices represent the three archetypes (GO, Des, and Asc). Points are color-coded according to their proximity to the archetypes. *Right*: Proportions of each archetype on the entire population: 37.8 % Asc (green), 39 % Des (blue) and 23.2 % GO (grey). **f** *Left*: Nicotine (Nic, pink triangles) and control (Ctl, grey dots) mice displayed on the same ternary plot. Nic mice displayed a visual shift of their behavior towards the GO extrema of the archetype. *Right*: This shift was reflected by a difference in the proportion of each phenotype between nicotine and control groups (χ^2 test, $p = 0.027$), with a higher proportion of GO mice in the nicotine group. **g** Archetypal composition for each archetype (1 = closer to the apex) in control and nicotine mice (Wilcoxon test, $p = 0.08$, $p = 0.22$ and $p = 0.04$, with Holm correction).

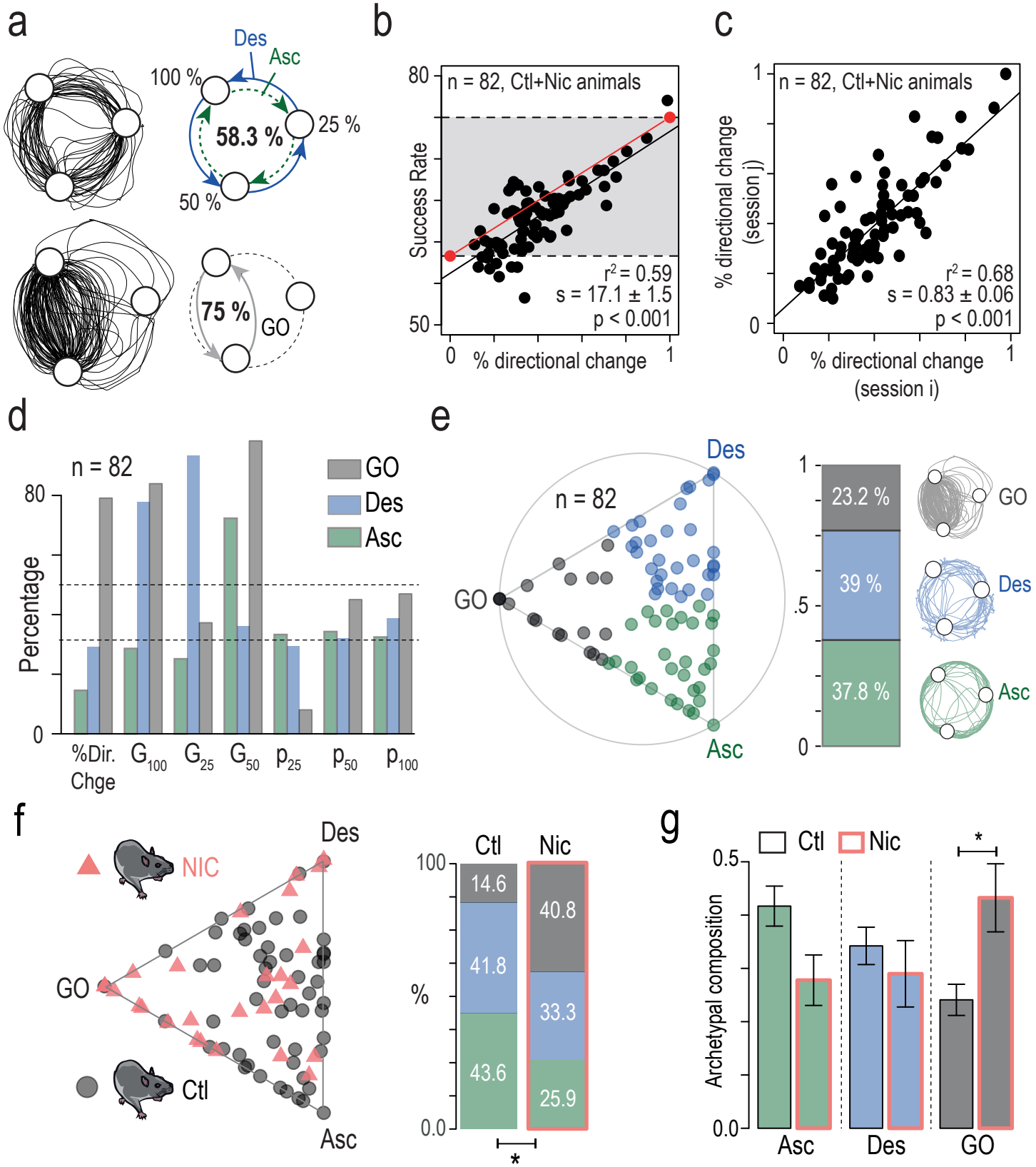

a
Des
Asc
100 %
25 %
58.3 %
50 %
75 %
GO
b
n = 82, Ctl+Nic animals
Success Rate
80
50
0
% directional change
1
r² = 0.59
s = 17.1 ± 1.5
p < 0.001
c
n = 82, Ctl+Nic animals
% directional change (session j)
1
0
% directional change (session i)
r² = 0.68
s = 0.83 ± 0.06
p < 0.001
d
n = 82
GO
Des
Asc
Percentage
80
0
%Dir. Chge
G100
G25
G50
p25
p50
p100
e
Des
n = 82
GO
Asc
1
.5
0
23.2 %
39 %
37.8 %
GO
Des
Asc
f
Des
NIC
GO
Ctl
Asc
Ctl
Nic
100
%
0.0
14.6
41.8
43.6
40.8
33.3
25.9
*
g
Ctl
Nic
0.5
0.0
Archetypal composition
*
Asc
Des
GO

Figure 3

Figure 4: Computational modeling suggests that decision parameters differ between the three archetypes and are differentially affected by nicotine exposure

a Principle of the softmax model: softmax decision rule with three parameters β (inverse temperature or exploration/exploitation), φ (uncertainty bonus) and κ (cost or effort for a directional change). **b** (Left) Estimated values of β, φ and κ parameters with M3 model for the 55 control mice (not exposed to nicotine, *** indicates a significant difference from zero). (Right) Bayesian Information Criterion (BIC) computed using three models of choice selection (Softmax with $\beta=0$, $\varphi=0$, $\kappa>0$ (M1), $\beta>0$, $\varphi=0$, $\kappa>0$ (M2) and $\beta>0$, $\varphi>0$, $\kappa>0$ (M3), see Methods). Smaller BIC value indicates that the uncertainty bonus provided a better fit. **c** Comparison between control data and model for a model sequence of 2000 choices (n = 55) simulated with fitted values of β, φ and κ (see b) *Left*: Repartition of visits on the three targets (p_{25}, p_{50} and p_{100}, with a mean of the differences between control data and model of Δ = 0.002%, -0.003% and 0.001%, p > 0.05). *Middle*: Comparison of the percentage of directional changes (Δ = 0.002, p > 0.05). *Right*: Probability to choose alternatives with the highest probability of reward for the three possible gambles (G_{100} = p_{50} over p_{25}; G_{25} = p_{100} over p_{50}; G_{50} = p_{100} over p_{25}, Δ = -0.02, -0.008 and 0.009, p > 0.05 for the three gambles). **d** Nicotine-exposed animals displayed an increase in β (Δ = 1.73, p = 0.036), but no difference in φ (Δ = 0.09, p = 0.29) and in κ (Δ = -0.07, p = 0.62) compared to control mice (Student's t-test with Holm correction for multiple comparisons). **e** Mimicking the effect of nicotine on the model parameters. *Left*: The simulation of choice behavior when nicotine-induced increase of β is added to the control model parameters (n = 55, Ctl + $\Delta\beta$) recapitulates the effect of nicotine on the three choice parameters (the probability to choose the most valuable option in gamble G_{100}; the percentage of directional changes, and the probability to visit p_{25}, mean of the differences between nicotine (Nic) data and model Δ = 0.007%, -0.009%, 0.003%, respectively, Student's t-test, p > 0.05). Starting from the nicotine mice parameters and removing the nicotine-induced changes on β (n = 27, Nic -$\Delta\beta$) reestablish those three parameters at the level of control mice (mean of the differences between control data and model Δ = -0.036%, -0.0006%, -0.02%, respectively, Student's t-test, p>0.05). $\Delta\beta$ is calculated using $\beta_{Nic} - \beta_{Ctl}$ the mean estimated in control and nicotine condition. **f** *Left:* Correlation between β (left), φ (middle) or κ (right) values and the archetypal composition for both control and nicotine mice (n = 82, see plot Figure 3b). The closer to the GO phenotype, the higher the β, which is consistent with an optimal strategy based on alternation between p_{100} and p_{50}. The closer to the Des phenotype, the lower the φ parameter.

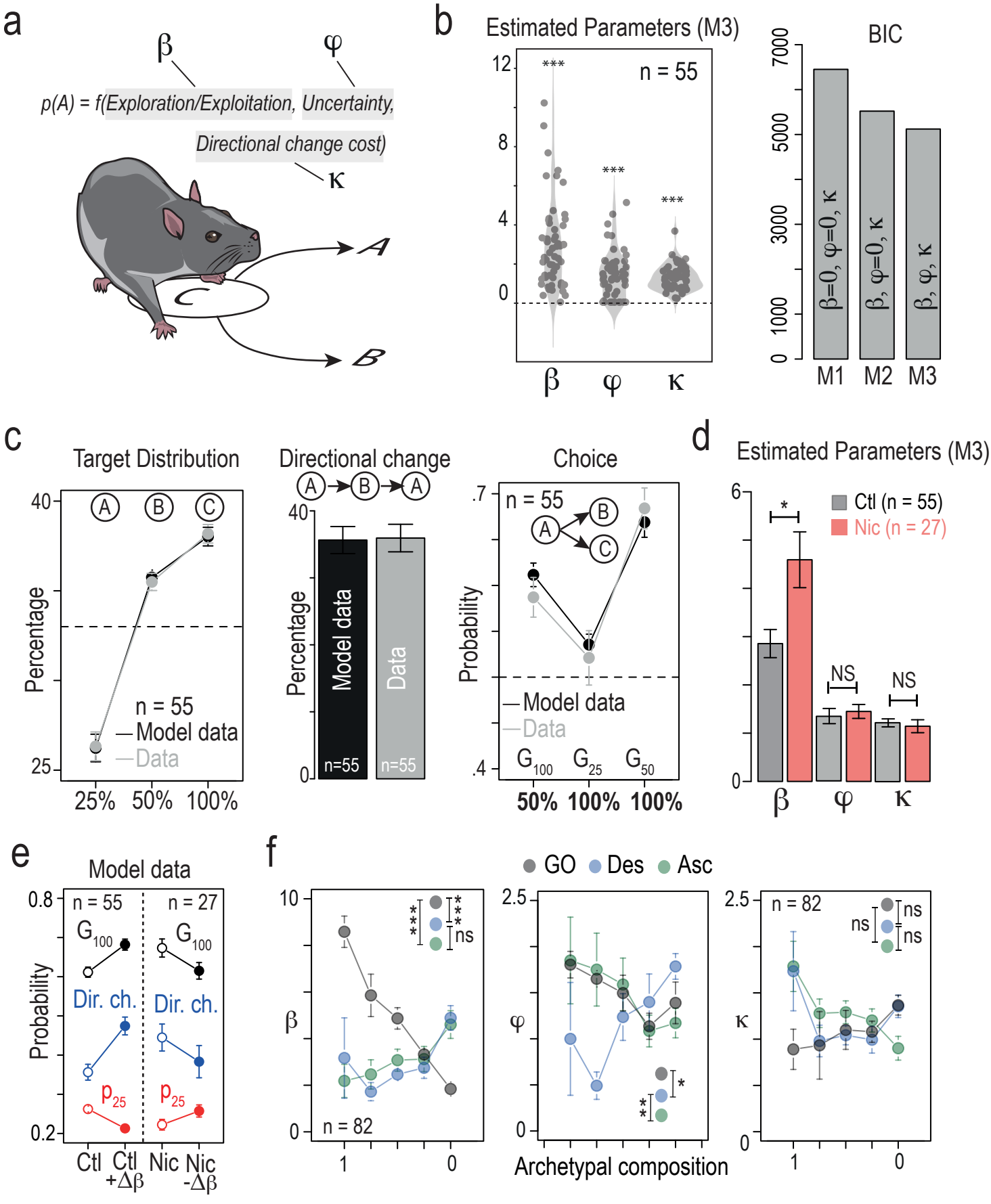

a
β
φ
p(A) = f(Exploration/Exploitation, Uncertainty,
Directional change cost)
κ
A
C
B
b
Estimated Parameters (M3)
n = 55
***
***
***
12
8
4
0
β
φ
κ
BIC
7000
5000
3000
1000
0
β=0, φ=0, κ
β, φ=0, κ
β, φ, κ
M1
M2
M3
c
Target Distribution
40
25
Percentage
A
B
C
n = 55
Model data
Data
25%
50%
100%
Directional change
A → B → A
40
0
Percentage
Model data
Data
n=55
n=55
Choice
.7
.4
Probability
n = 55
A
B
C
Model data
Data
G100
G25
G50
50%
100%
100%
d
Estimated Parameters (M3)
6
0
*
Ctl (n = 55)
Nic (n = 27)
NS
NS
β
φ
κ
e
Model data
0.8
0.2
Probability
n = 55
n = 27
G100
G100
Dir. ch.
Dir. ch.
p25
p25
Ctl
Ctl
+Δβ
Nic
Nic
-Δβ
f
GO
Des
Asc
10
0
β
***
**
ns
n = 82
1
0
2.5
0
φ
**
*
Archetypal composition
2.5
0
κ
n = 82
ns
ns
ns
1
0

Figure 4

Figure 5: Optogenetic manipulation of VTA DA neuron activity recapitulated the behavioral adaptations observed under chronic nicotine exposure.

a *Left:* representative ex vivo current-clamp recording of a VTA DA neuron transduced with CatCh and stimulated with 5-ms blue light pulses at 8 Hz. *Right*: Average increase in basal firing frequency upon optogenetic stimulation for n = 10 neurons (p-value = 0.002, Wilcoxon test). **b** *Top left:* representative ex vivo current-clamp recording of a VTA DA neuron transduced with Jaws and stimulated with 500-ms green light pulses at 0.5 Hz. *Bottom left:* Raster plot for n = 16 neurons. *Right*: Average decrease in basal firing frequency upon optogenetic stimulation, and return to the baseline after the photostimulation period, for n = 16 neurons (p-value: ns = 0.18; **0.004; **0.0014, Wilcoxon test with Holm correction). **c** Task design and photo-stimulation protocols. DAT^{iCRE} mice transduced with either an AAV-DIO-Catch-YFP in the VTA (CatCh, blue) or an AAV-DIO-Jaws-eGFP (Jaws, yellow) and were implanted unilaterally with bipolar stimulating electrodes for ICSS in the MFB. Following the deterministic (DS) and probabilistic setting (PS) sessions they received 2 paired Off (*filled circles*) and Off (*open circles*) sessions with the same rules as in the probabilistic setting. *Below:* Representative trajectories of a CatCh and Jaws-transduced mouse with (*blue and Yellow*) and without (*black*) optogenetic stimulation of VTA DA neurons. **d** (Left) Percentage of directional changes and (right) Target Distribution (p25) in individual mice (grey points) for CatCh (n=10) and Jaws-transduced (n=11) mice during On and Off sessions. Mean + sem are in black. Red arrows indicate the net effect of nicotine for comparison (Student's t-test or Wilcoxon test, unilateral testing). **e** Same as D for the softmax model β parameter. **f** *Left*: Position of each animal in the ternary archetype plot for CatCh (n=10, above) and Jaws-transduced (n=11, below) mice, in On (Blue or yellow points) and Off (Grey point) sessions. *Right*: (above) GO archetypal composition in individual mice (grey points) during On and Off session, for CatCh (n=10) and Jaws-transduced (n=11) mouse. Mean + sem are in black. (below) Net effect (ON-OFF) of light stimulation in archetypal composition for each archetype. Optogenetic activation of DA neurons triggered a shift of the behavior towards the GO phenotype while optogenetic inhibition induced a shift of the behavior away from GO (Student's t-test, unilateral testing).

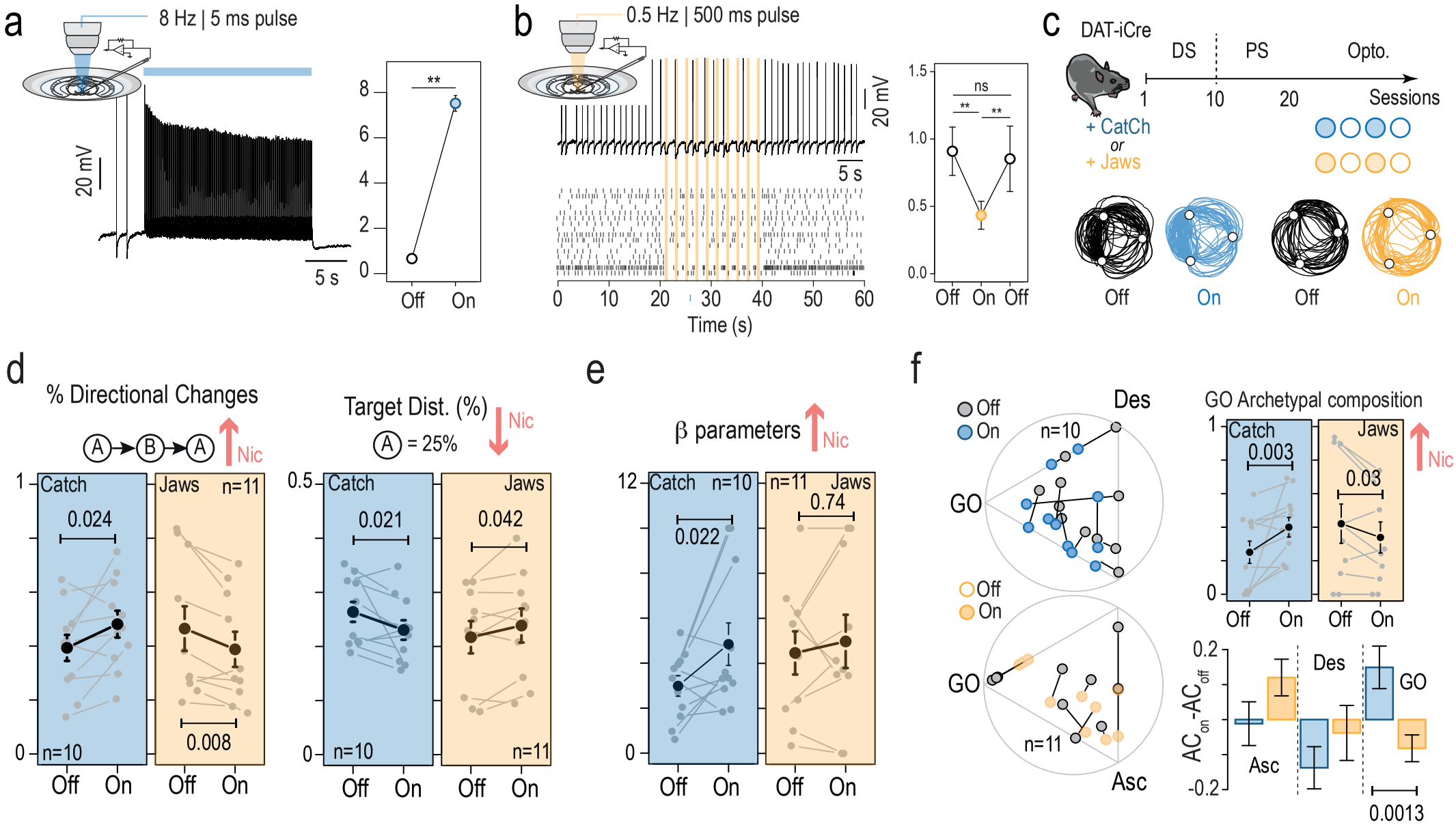
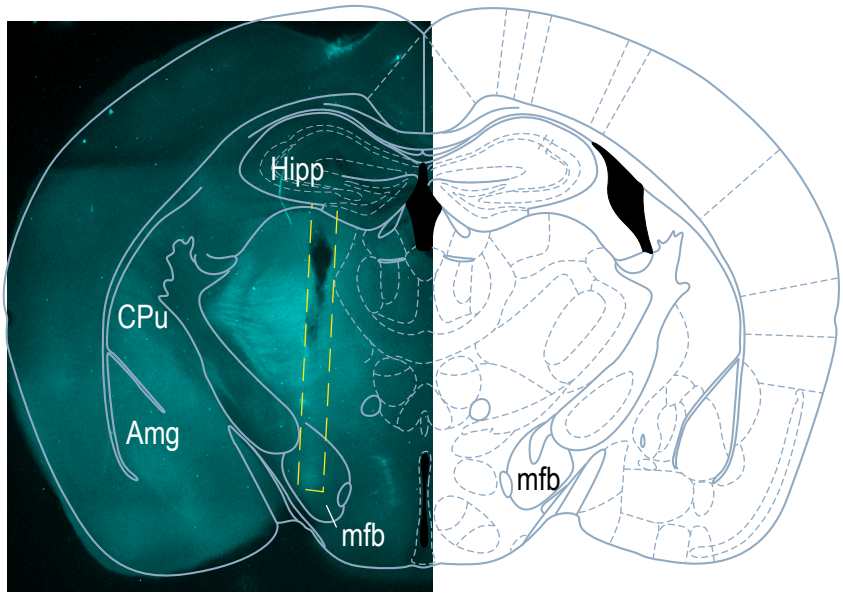

a
8 Hz | 5 ms pulse
20 mV
5 s
**
Off
On
b
0.5 Hz | 500 ms pulse
20 mV
5 s
Time (s)
ns
**
**
c
DAT-iCre
DS
PS
Opto.
Sessions
+ CatCh
or
+ Jaws
Off
On
Off
On
d
% Directional Changes
Nic
Target Dist. (%)
A = 25%
Nic
Catch
Jaws
n=11
0.024
0.008
n=10
0.021
0.042
e
β parameters
Nic
Catch
n=10
n=11
Jaws
0.022
0.74
f
Off
On
Des
n=10
GO
Off
On
GO
n=11
Asc
GO Archetypal composition
Catch
Jaws
0.003
0.03
Nic
Off
On
Off
On
AC_on-AC_off
Des
GO
Asc
0.0013

Figure 5

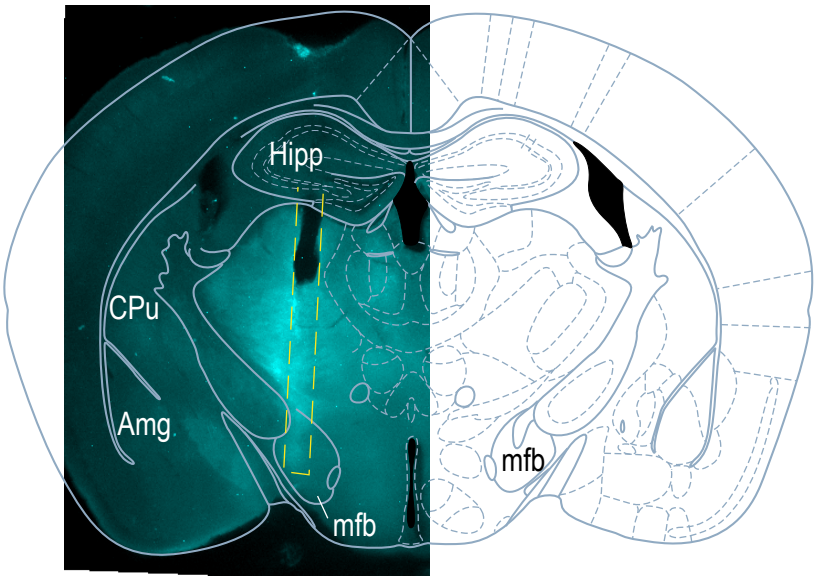
Supplementary Information

Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration. Malou Dongelmans, Romain Durand-de Cuttoli, Claire Nguyen, Maxime Come, Etienne K. Duranté, Damien Lemoine, Raphael Britto, Tarek Ahmed Yahia, Sarah Mondoloni, Steve Didienne, Elise Bousseyrol, Bernadette Hannesse, Lauren M. Reynolds, Nicolas Torquet, Deniz Dalkara, Fabio Marti, Alexandre Mourot, Jérémie Naudé, Philippe Faure

Supplementary Figure 1: **Stimulating electrode implantation:** Representative examples of unilateral MFB implantations in two different brains. Post-hoc verification of the ICSS track is represented in dotted yellow line.



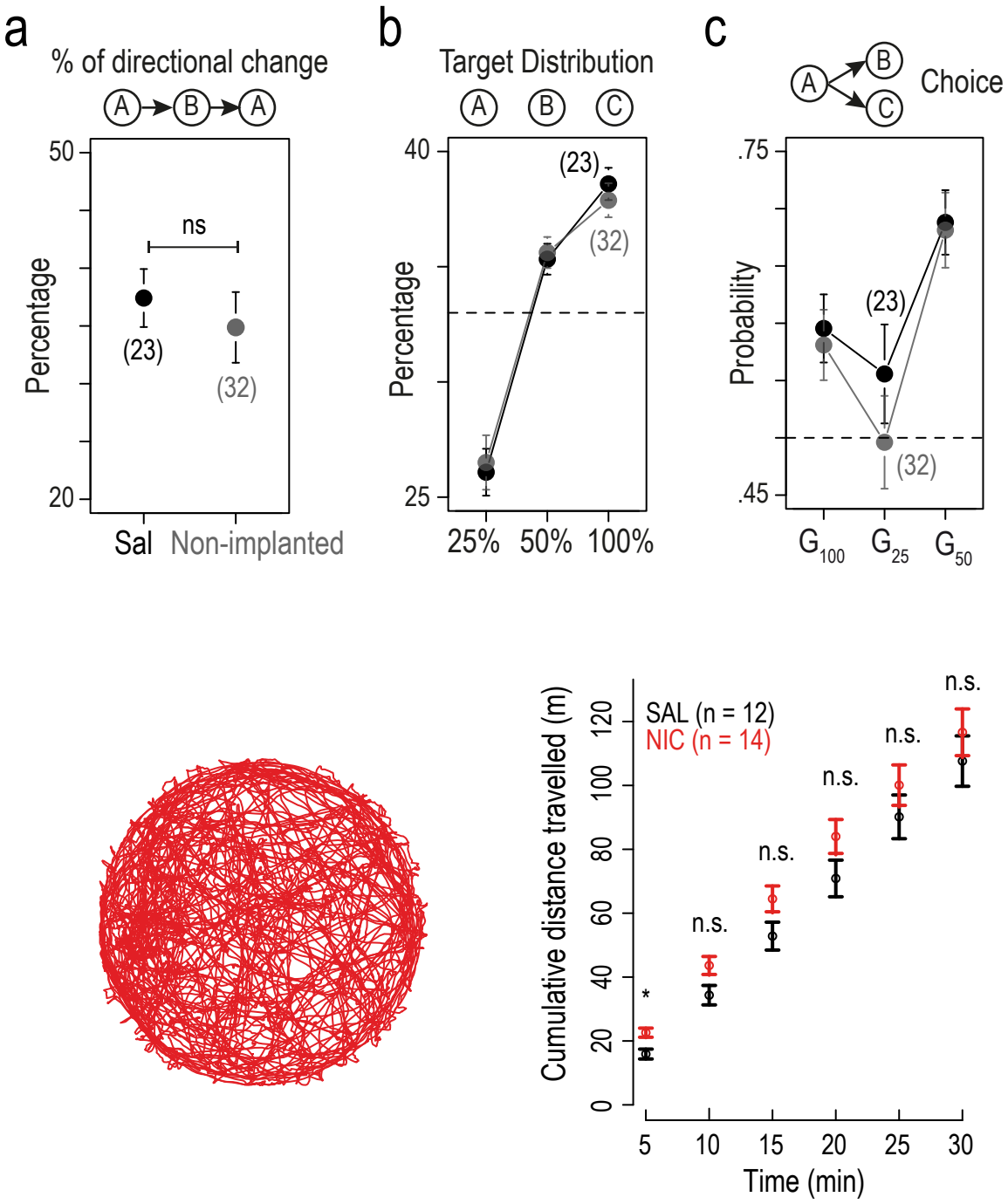
Bregma -1.34 mm
Yellow (ICSS)



Bregma -1.34 mm
Yellow (ICSS)

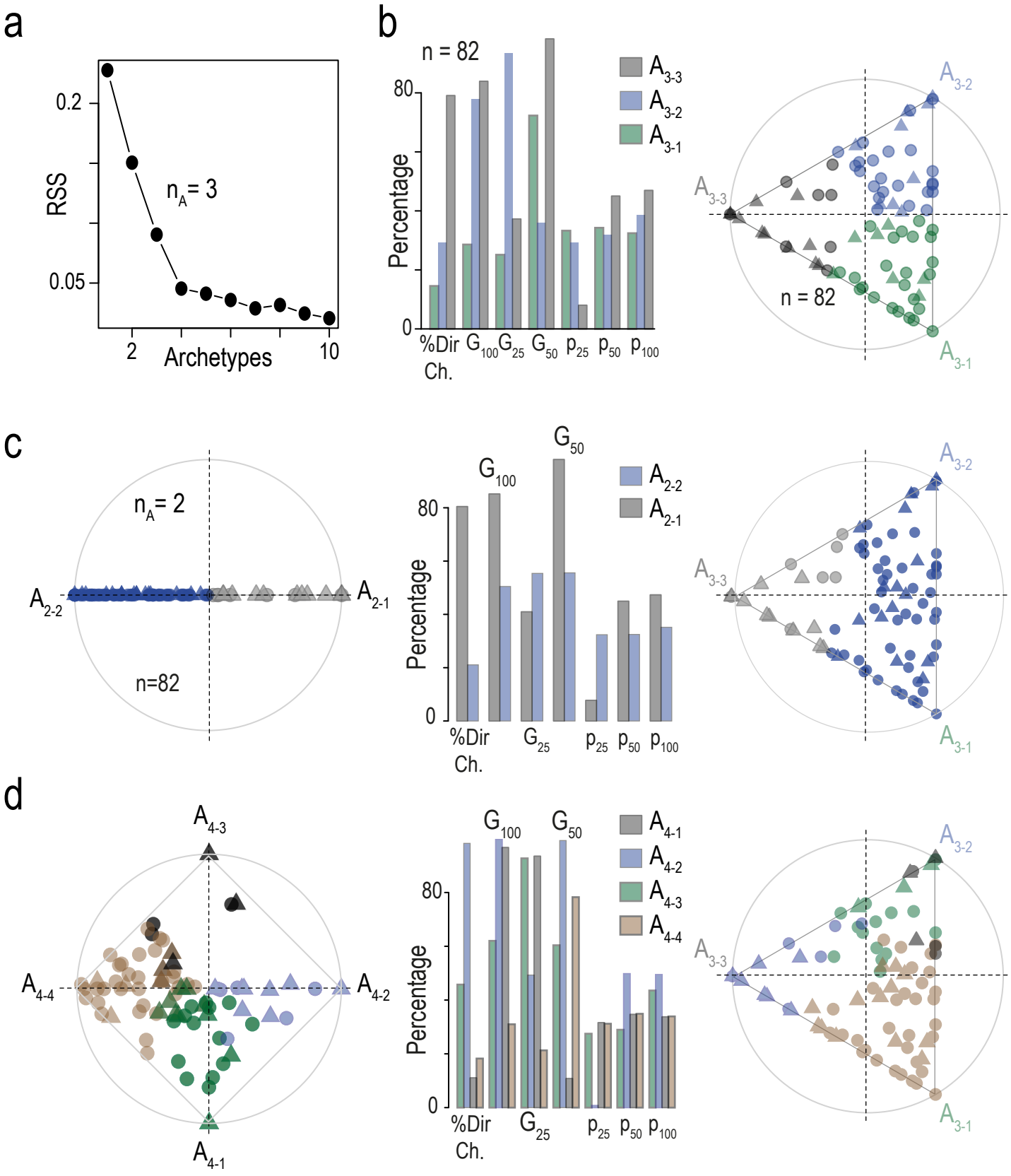
Supplementary Figure 1

Supplementary Figure 2: **a-c No behavioral difference between mice implanted with an osmotic minipump filled with saline (Sal) and non-implanted mice. a** Comparison of the percentage of directional change in Sal (black) and non-implanted (grey) mice (Wilcoxon signed rank test, $p > 0.05$). **b** Repartition of the visits to the three targets in the deterministic setting (DS, Wilcoxon Test $p > 0.05$ for the three comparisons). **c** Probability to choose the alternative with the highest probability of reward for the three possible gambles: G_{100}; G_{25} and G_{50} (Wilcoxon Test $p > 0.05$ for the three comparisons). **d Nicotine-treated mice show increased locomotion for the first five minutes in an open field.** Top, trajectory in an open field (duration 30 minutes) of a mouse treated for 24 days with nicotine (10 mg/kg/day). Bottom, cumulative distance travelled in meters measured every 5 minutes during a 30 min-OF exploration. Nic mice (n = 14) showed a greater distance travelled during the first 5 minutes only (t-test, $t = -2.4154$, $df = 22.074$, $p = 0.02444$), compared to saline-exposed mice (Sal, n = 12). The total distance travelled after 30 minutes was not significantly different between the two groups ($p > 0.05$).



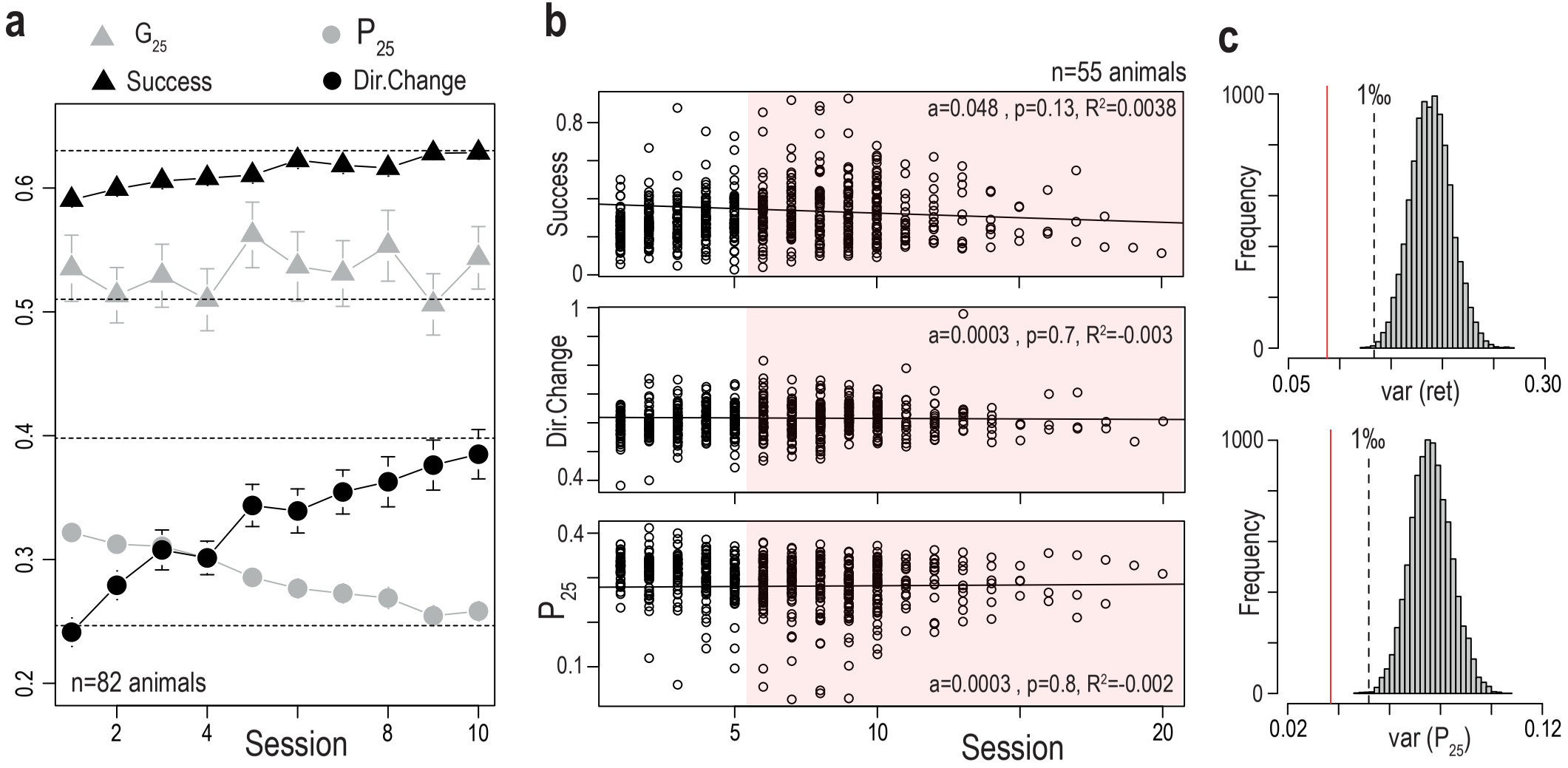
Supplementary figure 2

Supplementary Figure 3: **Archetypal analyses with 2 to 4 apices**: Archetypal analysis of the choice strategies based on a 7-dimensional data space: % of directional change, Gambles G_{100} (choice 50% over 25%), G_{25} (100% over 50%) and G_{50} (100% over 25%), and probabilities of choosing each point (P_{25}, P_{50}, and P_{100}). Analyses were performed on n = 82 mice (pooled control and nicotine mice). **a** Residual sum of squares for a number or archetypes n_A = 1 to 10. Error reduction between n_A = 4 and 10 is marginal. **b** Plot of the archetypal solutions for n_A = 3. Left: Percentile plot of the value of the 7 basic variables used in this analysis for the three archetypes, here labelled $A_{3\text{-}1}$ to $A_{3\text{-}3}$. $A_{3\text{-}1}$, $A_{3\text{-}2}$ and $A_{3\text{-}3}$ correspond to the GO, Des and Asc archetypes of Figure 3d, but are unlabeled here for comparison purposes with n_A = 2 and 4. Right: Visualization of the α coefficients using a ternary plot, in which the three apices represent the three archetypes. Each point shows the projection of each individual (n=82). Points are color-coded according to their proximity to the archetypes. **c** Plot of the archetypal solutions $A_{2\text{-}1}$ and $A_{2\text{-}2}$ for n_A = 2. From left to right: Visualization of the α coefficients in a binary plot, percentile plot and ternary plot (same as in B, with points color-coded according to their proximity to the two archetypes $A_{2\text{-}1}$ and $A_{2\text{-}2}$). **d** Same as C for n_A = 4 ($A_{4\text{-}1}$ to $A_{4\text{-}4}$). Note that the $A_{3\text{-}3}$ (a.k.a. GO) archetype is present when both n_A = 2 (the $A_{2\text{-}1}$ archetype) and n_A = 4 ($A_{4\text{-}2}$).



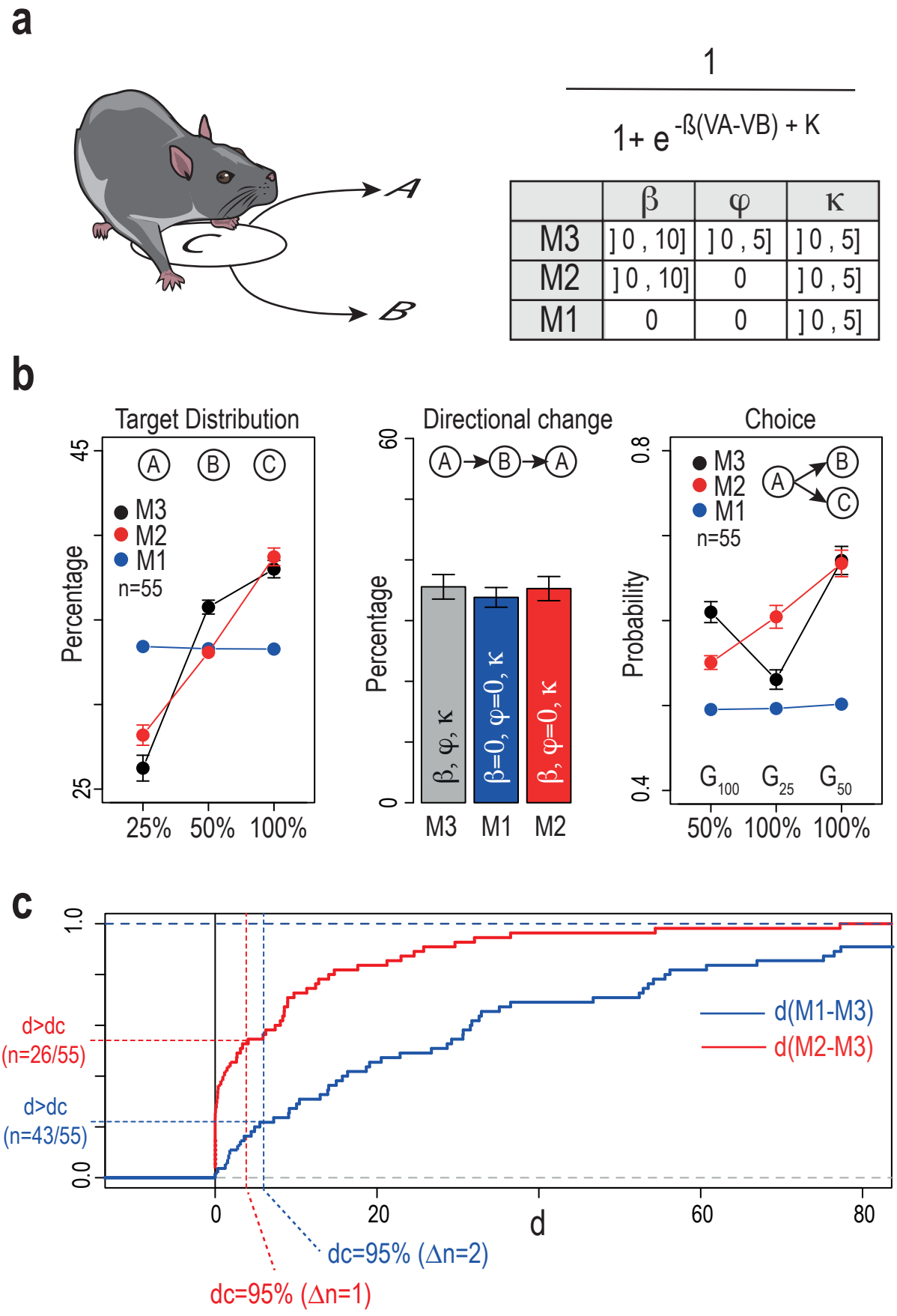
Supplementary Figure 3

Supplementary Figure 4: **Evolution of choice behavior parameters with session. a** Adaptation to the task in the probabilistic setting (PS) with convergence to the mean value (Dashed lines) after few sessions for the success rate, the probability to choose the option with the highest probability of reward in gamble G_{25}, the directional change and the percentage of visits to the 25% target (P_{25}). Mean values (dashed lines) were obtained by means of last two days of probabilistic setting for each individual (n=82, pooled control and nicotine-treated)) mice. **b** Individual values (each point represents one mouse in a given session) for the directional change (above), the success rate (Middle) and the percentage of visits to the 25% target (P_{25}). Black lines indicate linear regression model between session and individual values for session higher than 4 (red background). Estimated slope value (a), probability (p) to reject the null hypothesis of a null slope and the square of the correlation coefficient (R^2). This data indicates no change of value with session number after the first 5 sessions. **c** Distribution (n=10000 estimates) of the absolute value of the difference in directional change (above) and in the percentage of visits to the 25% target (P_{25}) for 2 sessions (higher than the fifth session) in different animals. Vertical dashed line indicated the 1‰ quantile. Vertical red line indicates the mean of the absolute value of the difference estimated between consecutive sessions within the same animals. Inter-individual variations are lower than the intra-individual variations.



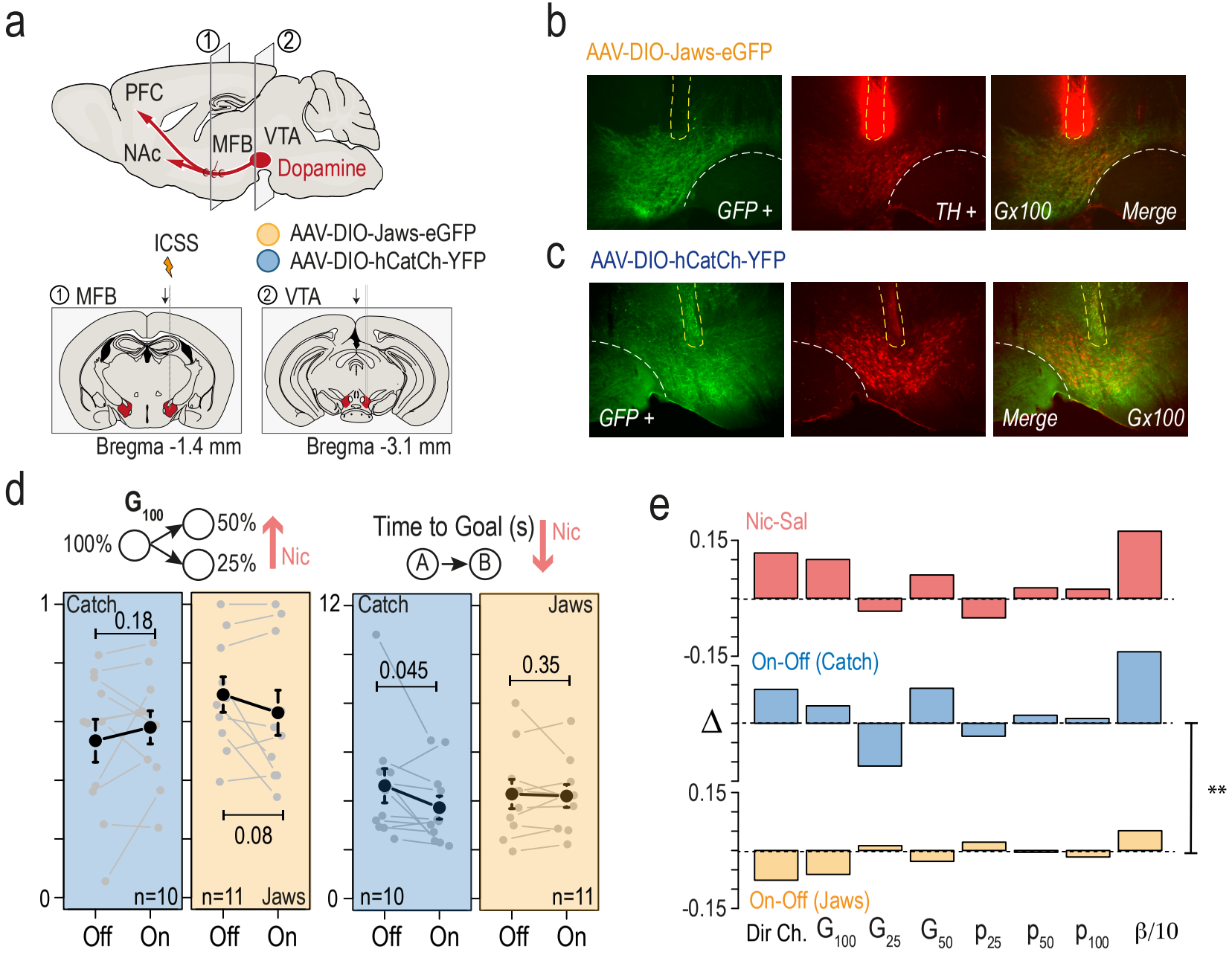
Supplementary Figure 4

Supplementary Figure 5: **Principle of the softmax model: a** softmax decision rule with three parameters β (inverse temperature or exploration/exploitation), φ (uncertainty bonus) and κ (cost or effort for a directional change). Three models can be derived from the main model labelled M3 for which ($\beta>0$, $\varphi>0$, $\kappa>0$) M1 and M2 are two special cases of M3. M1: $\beta=0$, $\varphi=0$, $\kappa>0$, M2: $\beta>0$, $\varphi=0$, $\kappa>0$ (see Methods). The table indicates the range of values for the three parameters depending on the model M1, M2 or M3. **b** Model sequence of 2000 choices (n = 55) simulated with fitted values of β, φ and κ for M1, M2 and M3 (same data as in Figure 4c for M3) Left: Repartition of visits on the three targets (p_{25}, p_{50} and p_{100}), Middle: percentage of directional changes and Right: Probability to choose alternatives with the highest probability of reward for the three possible gambles (G_{100} = p_{50} over p_{25}; G_{25} = p_{100} over p_{50}; G_{50} = p_{100} over p_{25}). **c** Density of twice the difference in log likelihoods of M1 and M3 (blue line) and of M1 and M3 (red line). Vertical dashed line indicates critical d value (d_c) for 95% significance of ratio test (see Methods). Colored horizontal line indicates number of d>dc for each comparison.



Supplementary Figure 5

Supplementary Figure 6: **Optogenetic manipulation of choice behaviors. a** DAT^{iCRE} mice were implanted unilaterally with bipolar stimulating electrodes for ICSS in the medial forebrain bundle (1) and transduced with either an AAV-DIO-Jaws-eGFP or an AAV-DIO-Catch-YFP in the VTA (2). **b** Representative immunohistochemical verification of CatCh-YFP expression selectively in DA neurons of the VTA (anti-TH in red, and -GFP in green; merge on top). Post-hoc verification of the unilateral fiber implantation is represented by dotted yellow lines. **c** Representative immunohistochemical verification of Jaws-eGFP expression selectively in DA neurons of the VTA (anti-TH in red, and -GFP in green; merge on top). Post-hoc verification of the unilateral fiber implantation is represented by dotted yellow lines. **d** (Left) Percent of choice toward P_{50} in gamble G_{100} and (Right) the time to goal in individual mice (grey points) for CatCh (n=10) and Jaws-transduced (n=11) mice during On and Off sessions. Mean + sem are in black. Red arrows indicate the net effect of nicotine for comparison (Student's t-test or Wilcoxon test, unilateral testing). **e** Net effect of light stimulation on eight parameters (b values are divided by ten for presentation). Data from OFF sessions were subtracted from data from ON sessions. In red, the effect of nicotine (NIC) is represented for all parameters, as a comparison factor. (Asterisk: chi-squared test to compare the number of parameters that vary in the same direction as nicotine during photostimulation or photoinhibition).



Supplementary figure 6

Bibliographie

Addicott, M.A., Pearson, J.M., Sweitzer, M.M., Barack, D.L., Platt, M.L., 2017. A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. Neuropsychopharmacol 42, 1931–1939.

Alvernhe, A., Save, E., Poucet, B., 2011. Local remapping of place cell firing in the Tolman detour task. European Journal of Neuroscience 33, 1696–1705.

Amarante, L.M., Laubach, M., 2021. Coherent theta activity in the medial and orbital frontal cortices encodes reward value. eLife 10, e63372.

Andén, N.-E., Carlsson, A., Dahlström, A., Fuxe, K., Hillarp, N.-Å., Larsson, K., 1964. Demonstration and mapping out of nigro-neostriatal dopamine neurons. Life Sciences 3, 523–530.

Anselme, P., Robinson, M.J.F., Berridge, K.C., 2013. Reward uncertainty enhances incentive salience attribution as sign-tracking. Behavioural Brain Research 238, 53–61.

Bach, D.R., Dolan, R.J., 2012. Knowing how much you don't know: a neural organization of uncertainty estimates. Nat Rev Neurosci 13, 572–586.

Balleine, B.W., 2019. The Meaning of Behavior: Discriminating Reflex and Volition in the Brain. Neuron 104, 47–62.

Banich, M.T., Floresco, S., 2019. Reward systems, cognition, and emotion: Introduction to the special issue. Cogn Affect Behav Neurosci 19, 409–414.

Barik, J., Marti, F., Morel, C., Fernandez, S.P., Lanteri, C., Godeheu, G., Tassin, J.-P., Mombereau, C., Faure, P., Tronche, F., 2013. Chronic Stress Triggers Social Aversion via Glucocorticoid Receptor in Dopaminoceptive Neurons. Science 339, 332–335.

Barter, J.W., Li, S., Lu, D., Bartholomew, R.A., Rossi, M.A., Shoemaker, C.T., Salas-Meza, D., Gaidis, E., Yin, H.H., 2015. Beyond reward prediction errors: the role of dopamine in movement kinematics. Frontiers in Integrative Neuroscience 9, 39.

Bayer, H.M., Glimcher, P.W., 2005. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron 47, 129–141.

Bechara, A., Damasio, H., Damasio, A.R., 2000. Emotion, Decision Making and the Orbitofrontal Cortex. Cerebral Cortex 10, 295–307.

Beeler, J.A., 2012. Thorndike's Law 2.0: Dopamine and the Regulation of Thrift. Front. Neurosci. 6.

Beier, K.T., Gao, X.J., Xie, S., DeLoach, K.E., Malenka, R.C., Luo, L., 2019. Topological Organization of Ventral Tegmental Area Connectivity Revealed by Viral-Genetic Dissection of Input-Output Relations. Cell Reports 26, 159-167.e6.

Beier, K.T., Steinberg, E.E., DeLoach, K.E., Xie, S., Miyamichi, K., Schwarz, L., Gao, X.J., Kremer, E.J., Malenka, R.C., Luo, L., 2015. Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. Cell 162, 622–634.

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P.L., Gioanni, Y., Battaglia, F.P., Wiener, S.I., 2010. Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal- Prefrontal Network upon Learning. Neuron 66, 921–936.

Berke, J.D., 2018. What does dopamine mean? Nature Neuroscience 21, 787–793.

Berlyne, D.E., 1966. Curiosity and Exploration. Science 153, 25–33.

Berridge, K.C., 1999. Pleasure, pain, desire, and dread: Hidden core processes of emotion. In: Well-Being: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York, NY, US, pp. 525–557.

Berridge, K.C., 2004. Motivation concepts in behavioral neuroscience. Physiology & Behavior, Reviews on Ingestive Science 81, 179–209.

Berridge, K.C., 2007. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191, 391–431.
Berridge, K.C., 2012. From prediction error to incentive salience: mesolimbic computation of reward motivation: From prediction error to incentive salience. European Journal of Neuroscience 35, 1124–1143.
Berridge, K.C., Robinson, T.E., 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research Reviews 28, 309–369.
Berridge, K.C., Robinson, T.E., 2003. Parsing reward. Trends in Neurosciences 26, 507–513.
Berridge, K.C., Robinson, T.E., Aldridge, J.W., 2009. Dissecting components of reward: 'liking', 'wanting', and learning. Current Opinion in Pharmacology, Neurosciences 9, 65–73.
Bilder, R.M., Volavka, J., Lachman, H.M., Grace, A.A., 2004. The Catechol-O-Methyltransferase Polymorphism: Relations to the Tonic–Phasic Dopamine Hypothesis and Neuropsychiatric Phenotypes. Neuropsychopharmacol 29, 1943–1961.
Bindra, D., 1974. A motivational view of learning, performance, and behavior modification. Psychological Review 81, 199–213.
Björklund, A., Dunnett, S.B., 2007. Dopamine neuron systems in the brain: an update. Trends in Neurosciences, Fifty years of dopamine research 30, 194–202.
Blanchard, T.C., Hayden, B.Y., Bromberg-Martin, E.S., 2015. Orbitofrontal Cortex Uses Distinct Codes for Different Choice Attributes in Decisions Motivated by Curiosity. Neuron 85, 602–614.
Bogacz, R., 2020. Dopamine role in learning and action inference. eLife 9, e53262.
Bolles, R.C., 1972. Reinforcement, expectancy, and learning. Psychological Review 79, 394–409.
Bromberg-Martin, E.S., Hikosaka, O., 2009. Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards. Neuron 63, 119–126.
Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O., 2010. Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. Neuron 68, 815–834.
Bryden, D.W., Roesch, M.R., 2015. Executive Control Signals in Orbitofrontal Cortex during Response Inhibition. Journal of Neuroscience 35, 3903–3914.
Burgers, J.M., 1966. Curiosity and Play: Basic Factors in the Development of Life. Science 154, 1680–1681.
Buzsáki, G., 2002. Theta Oscillations in the Hippocampus. Neuron 33, 325–340.
Buzsaki, G., 2004. Neuronal Oscillations in Cortical Networks. Science 304, 1926–1929.
Buzsáki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nature Reviews Neuroscience 13, 407–420.
Buzsáki, G., Logothetis, N., Singer, W., 2013. Scaling Brain Size, Keeping Timing: Evolutionary Preservation of Brain Rhythms. Neuron 80, 751–764.
Cai, X., Padoa-Schioppa, C., 2014. Contributions of Orbitofrontal and Lateral Prefrontal Cortices to Economic Choice and the Good-to-Action Transformation. Neuron 81, 1140–1151.
Carlén, M., 2017. What constitutes the prefrontal cortex? Science 358, 478–482.
Cavanagh, J.F., Frank, M.J., 2014. Frontal theta as a mechanism for cognitive control. Trends in Cognitive Sciences 18, 414–421.
Chiara, G.D., Imperato, A., 1988. Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. PNAS 85, 5274–5278.
Cisek, P., 2007. Cortical mechanisms of action selection: the affordance competition hypothesis. Philosophical Transactions of the Royal Society B: Biological Sciences 362, 1585–1599.

Cisek, P., 2012. Making decisions through a distributed consensus. Current Opinion in Neurobiology, Decision making 22, 927–936.
Coddington, L.T., Dudman, J.T., 2018. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci 21, 1563–1573.
Cohen, J.D., McClure, S.M., Yu, A.J., 2007. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society B: Biological Sciences 362, 933–942.
Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., Uchida, N., 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482, 85–88.
Constantinople, C.M., Piet, A.T., Bibawi, P., Akrami, A., Kopec, C., Brody, C.D., 2019. Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases. eLife 8, e49744.
Cools, R., 2019. Chemistry of the Adaptive Mind: Lessons from Dopamine. Neuron 104, 113–131.
Costa, V.D., Tran, V.L., Turchi, J., Averbeck, B.B., 2014. Dopamine modulates novelty seeking behavior during decision making. Behavioral Neuroscience 128, 556.
Craig, W., 1917. Appetites and Aversions as Constituents of Instincts. Proc Natl Acad Sci U S A 3, 685–688.
da Silva, J.A., Tecuapetla, F., Paixão, V., Costa, R.M., 2018. Dopamine neuron activity before action initiation gates and invigorates future movements. Nature 554, 244–248.
Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., Botvinick, M., 2020. A distributional code for value in dopamine-based reinforcement learning. Nature 577, 671–675.
Dautan, D., Souza, A.S., Huerta-Ocampo, I., Valencia, M., Assous, M., Witten, I.B., Deisseroth, K., Tepper, J.M., Bolam, J.P., Gerdjikov, T.V., Mena-Segovia, J., 2016. Segregated cholinergic transmission modulates dopamine neurons integrated in distinct functional circuits. Nat Neurosci 19, 1025–1033.
Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience 8, 1704–1711.
Day, J.J., Roitman, M.F., Wightman, R.M., Carelli, R.M., 2007. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci 10, 1020–1028.
Dayan, P., 2009. Goal-directed control and its antipodes. Neural Networks, Goal-Directed Neural Systems 22, 213–219.
Dayan, P., Berridge, K.C., 2014. Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. Cogn Affect Behav Neurosci 14, 473–492.
Dayan, P., Niv, Y., 2008. Reinforcement learning: The Good, The Bad and The Ugly. Current Opinion in Neurobiology, Cognitive neuroscience 18, 185–196.
Dayan, P., Niv, Y., Seymour, B., Daw, N.D., 2006. The misbehavior of value and the discipline of the will. Neural Networks 8.
de Jong, J.W., Afjei, S.A., Pollak Dorocic, I., Peck, J.R., Liu, C., Kim, C.K., Tian, L., Deisseroth, K., Lammel, S., 2019. A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. Neuron 101, 133-151.e7.
Descarries, L., Bérubé-Carrière, N., Riad, M., Bo, G.D., Mendez, J.A., Trudeau, L.-É., 2008. Glutamate in dopamine neurons: Synaptic versus diffuse transmission. Brain Research Reviews, Basal Ganglia 58, 290–302.
Di Chiara, G., 1995. The role of dopamine in drug abuse viewed from the perspective of its role in motivation. Drug and Alcohol Dependence, Commonalities in mechanisms of drug dependence 38, 95–137.
Dickinson, A., Balleine, B., 1994. Motivational control of goal-directed action. Animal Learning & Behavior 22, 1–18.

Dolan, R.J., Dayan, P., 2013. Goals and Habits in the Brain. Neuron 80, 312–325.
Drai, D., Benjamini, Y., Golani, I., 2000. Statistical discrimination of natural modes of motion in rat exploratory behavior. Journal of Neuroscience Methods 96, 119–131.
Durstewitz, D., Seamans, J.K., 2008. The Dual-State Theory of Prefrontal Cortex Dopamine Function with Relevance to Catechol-O-Methyltransferase Genotypes and Schizophrenia. Biological Psychiatry 64, 739–749.
Durstewitz, D., Vittoz, N.M., Floresco, S.B., Seamans, J.K., 2010. Abrupt Transitions between Prefrontal Neural Ensemble States Accompany Behavioral Transitions during Rule Learning. Neuron 66, 438–448.
Dvorkin, A., Benjamini, Y., Golani, I., 2008. Mouse Cognition-Related Behavior in the Open-Field: Emergence of Places of Attraction. PLOS Computational Biology 4, e1000027.
Edwards, W., 1954. The Theory of Decision Making. Psychological Bulletin 51, 38.
Eilam, D., Golani, I., 1989. Home base behavior of rats (Rattus norvegicus) exploring a novel environment. Behavioural Brain Research 34, 199–211.
Eshel, N., Tian, J., Bukwich, M., Uchida, N., 2016. Dopamine neurons share common response function for reward prediction error. Nat Neurosci 19, 479–486.
Estes, W.K., 1943. Discriminative conditioning. I. A discriminative property of conditioned anticipation. Journal of Experimental Psychology 32, 150–155.
Euston, D.R., Gruber, A.J., McNaughton, B.L., 2012. The Role of Medial Prefrontal Cortex in Memory and Decision Making. Neuron 76, 1057–1070.
Faure, P., Tolu, S., Valverde, S., Naudé, J., 2014. Role of nicotinic acetylcholine receptors in regulating dopamine neuron activity. Neuroscience 282, 86–100.
Finlayson, K., Lampe, J.F., Hintze, S., Würbel, H., Melotti, L., 2016. Facial Indicators of Positive Emotions in Rats. PLOS ONE 11, e0166446.
Fiorillo, C.D., 2011. Transient activation of midbrain dopamine neurons by reward risk. Neuroscience 197, 162–171.
Fiorillo, C.D., Tobler, P.N., Schultz, W., 2003. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. Science, New Series 299, 1898–1902.
Fischbach-Weiss, S., Reese, R.M., Janak, P.H., 2018. Inhibiting Mesolimbic Dopamine Neurons Reduces the Initiation and Maintenance of Instrumental Responding. Neuroscience 372, 306–315.
Floresco, S.B., 2013. Prefrontal dopamine and behavioral flexibility: shifting from an "inverted-U" toward a family of functions. Frontiers in Neuroscience 7.
Floresco, S.B., Todd, C.L., Grace, A.A., 2001. Glutamatergic Afferents from the Hippocampus to the Nucleus Accumbens Regulate Activity of Ventral Tegmental Area Dopamine Neurons. J. Neurosci. 21, 4915–4922.
Floresco, S.B., West, A.R., Ash, B., Moore, H., Grace, A.A., 2003. Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. Nat Neurosci 6, 968–973.
Fonio, E., Benjamini, Y., Golani, I., 2009. Freedom of movement and the stability of its unfolding in free exploration of mice. PNAS 106, 21335–21340.
Fujisawa, S., Buzsáki, G., 2011. A 4 Hz Oscillation Adaptively Synchronizes Prefrontal, VTA, and Hippocampal Activities. Neuron 72, 153–165.
Gardner, M.P.H., Conroy, J.S., Shaham, M.H., Styer, C.V., Schoenbaum, G., 2017. Lateral Orbitofrontal Inactivation Dissociates Devaluation-Sensitive Behavior and Economic Choice. Neuron 96, 1192-1203.e4.
Gardner, M.P.H., Schoenbaum, G., 2021. The orbitofrontal cartographer. Behavioral Neuroscience 135, 267–276.
Gardner, M.P.H., Schoenbaum, G., Gershman, S.J., 2018. Rethinking dopamine as generalized prediction error. Proceedings of the Royal Society Biological Sciences 285.

Geisler, S., Derst, C., Veh, R.W., Zahm, D.S., 2007. Glutamatergic Afferents of the Ventral Tegmental Area in the Rat. Journal of Neuroscience 27, 5730–5743.
Gershman, S.J., Uchida, N., 2019. Believing in dopamine. Nat Rev Neurosci 20, 703–714.
Girault, J.-A., Greengard, P., 2004. The Neurobiology of Dopamine Signaling. Archives of Neurology 61, 641–644.
Glangetas, C., Fois, G.R., Jalabert, M., Lecca, S., Valentinova, K., Meye, F.J., Diana, M., Faure, P., Mameli, M., Caille, S., Georges, F., 2015. Ventral Subiculum Stimulation Promotes Persistent Hyperactivity of Dopamine Neurons and Facilitates Behavioral Effects of Cocaine. Cell Reports 13, 2287–2296.
Glimcher, P.W., 2009. Chapter 32 - Choice: Towards a Standard Back-pocket Model. In: Glimcher, P.W., Camerer, C.F., Fehr, E., Poldrack, R.A. (Eds.), Neuroeconomics, Decision Making and the Brain. Academic Press, London, pp. 503–521.
Gonzalez-Burgos, G., Lewis, D.A., 2008. GABA Neurons and the Mechanisms of Network Oscillations: Implications for Understanding Cortical Dysfunction in Schizophrenia. Schizophrenia Bulletin 34, 944–961.
Grace, A.A., Bunney, B.S., 1984a. The control of firing pattern in nigral dopamine neurons: single spike firing. J. Neurosci. 4, 2866–2876.
Grace, A.A., Bunney, B.S., 1984b. The control of firing pattern in nigral dopamine neurons: burst firing. J. Neurosci. 4, 2877–2890.
Granon, S., Changeux, J.-P., 2012. Deciding between conflicting motivations: What mice make of their prefrontal cortex. Behavioural Brain Research 229, 419–426.
Haber, S.N., Fudge, J.L., McFarland, N.R., 2000. Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum. J. Neurosci. 20, 2369–2382.
Hart, A.S., Clark, J.J., Phillips, P.E.M., 2015. Dynamic shaping of dopamine signals during probabilistic Pavlovian conditioning. Neurobiology of Learning and Memory 117, 84–92.
Hart, A.S., Rutledge, R.B., Glimcher, P.W., Phillips, P.E.M., 2014. Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term. J. Neurosci. 34, 698–704.
Heilbronner, S.R., Hayden, B.Y., Platt, M.L., 2010. Neuroeconomics of risk-sensitive decision making. In: Madden, G.J., Bickel, W.K. (Eds.), Impulsivity: The Behavioral and Neurological Science of Discounting. American Psychological Association, Washington, pp. 159–187.
Helfrich, R.F., Knight, R.T., 2016. Oscillatory Dynamics of Prefrontal Cognitive Control. Trends in Cognitive Sciences 20, 916–930.
Herrnstein, R.J., 1961. Relative and absolute strength of response as a function of frequency of reinforcement,. J Exp Anal Behav 4, 267–272.
Hnasko, T.S., Hjelmstad, G.O., Fields, H.L., Edwards, R.H., 2012. Ventral Tegmental Area Glutamate Neurons: Electrophysiological Properties and Projections. J. Neurosci. 32, 15076–15085.
Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E.M., Graybiel, A.M., 2013. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature 500, 575–579.
Hu, H., Cui, Y., Yang, Y., 2020. Circuits and functions of the lateral habenula in health and in disease. Nat Rev Neurosci 21, 277–295.
Hunt, L.T., Hayden, B.Y., 2017. A distributed, hierarchical and recurrent framework for reward-based choice. Nature Reviews Neuroscience 18, 172–182.
Ikemoto, S., 2007. Dopamine reward circuitry: Two projection systems from the ventral midbrain to the nucleus accumbens–olfactory tubercle complex. Brain Research Reviews 56, 27–78.

Isaacson, J.S., Scanziani, M., 2011. How Inhibition Shapes Cortical Activity. Neuron 72, 231–243.
Izquierdo, A., Suda, R.K., Murray, E.A., 2004. Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency. J. Neurosci. 24, 7540–7548.
Jenkinson, N., Brown, P., 2011. New insights into the relationship between dopamine, beta oscillations and motor function. Trends in Neurosciences 34, 611–618.
Jenni, N.L., Larkin, J.D., Floresco, S.B., 2017. Prefrontal Dopamine D1 and D2 Receptors Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and Amygdalar Circuits. J. Neurosci. 37, 6200–6213.
Jensen, P., Toates, F.M., 1993. Who needs 'behavioural needs'? Motivational aspects of the needs of animals. Applied Animal Behaviour Science 37, 161–181.
Jhou, T.C., Fields, H.L., Baxter, M.G., Saper, C.B., Holland, P.C., 2009. The Rostromedial Tegmental Nucleus (RMTg), a GABAergic Afferent to Midbrain Dopamine Neurons, Encodes Aversive Stimuli and Inhibits Motor Responses. Neuron 61, 786–800.
Jo, Y.S., Mizumori, S.J.Y., 2016. Prefrontal Regulation of Neuronal Activity in the Ventral Tegmental Area. Cereb Cortex 26, 4057–4068.
Juarez, B., Morel, C., Ku, S.M., Liu, Y., Zhang, H., Montgomery, S., Gregoire, H., Ribeiro, E., Crumiller, M., Roman-Ortiz, C., Walsh, J.J., Jackson, K., Croote, D.E., Zhu, Y., Zhang, S., Vendruscolo, L.F., Edwards, S., Roberts, A., Hodes, G.E., Lu, Y., Calipari, E.S., Chaudhury, D., Friedman, A.K., Han, M.-H., 2017. Midbrain circuit regulation of individual alcohol drinking behaviors in mice. Nat Commun 8, 2220.
Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10, 1625–1633.
Kakade, S., Dayan, P., 2002. Dopamine: generalization and bonuses. Neural Networks 15, 549–559.
Kebabian, J.W., Greengard, P., 1971. Dopamine-Sensitive Adenyl Cyclase: Possible Role in Synaptic Transmission. Science 174, 1346–1349.
Kepecs, A., Uchida, N., Zariwala, H.A., Mainen, Z.F., 2008. Neural correlates, computation and behavioural impact of decision confidence. Nature 455, 227–231.
Keramati, M., Gutkin, B., 2014. Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife 3, e04811.
Khamassi, M., Lallée, S., Enel, P., Procyk, E., Dominey, P., 2011. Robot Cognitive Control with a Neurophysiologically Inspired Reinforcement Learning Model. Frontiers in Neurorobotics 5, 1.
Kobayashi, S., Schultz, W., 2008. Influence of Reward Delays on Responses of Dopamine Neurons. Journal of Neuroscience 28, 7837–7846.
Lak, A., Okun, M., Moss, M.M., Gurnani, H., Farrell, K., Wells, M.J., Reddy, C.B., Kepecs, A., Harris, K.D., Carandini, M., 2019. Dopaminergic and Prefrontal Basis of Learning from Sensory Confidence and Reward Value. Neuron.
Lak, A., Stauffer, W.R., Schultz, W., 2014. Dopamine prediction error responses integrate subjective value from different reward dimensions. Proc Natl Acad Sci USA 111, 2343–2348.
Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B., Roeper, J., 2008. Unique Properties of Mesoprefrontal Neurons within a Dual Mesocorticolimbic Dopamine System. Neuron 57, 760–773.
Lammel, S., Lim, B.K., Ran, C., Huang, K.W., Betley, M.J., Tye, K.M., Deisseroth, K., Malenka, R.C., 2012. Input-specific control of reward and aversion in the ventral tegmental area. Nature 491, 212–217.

Le Van Quyen, M., Bragin, A., 2007. Analysis of dynamic brain oscillations: methodological advances. Trends in Neurosciences, July INMED/TINS special issue—Physiogenic and pathogenic oscillations: the beauty and the beast 30, 365–373.
Lebreton, M., Jorge, S., Michel, V., Thirion, B., Pessiglione, M., 2009. An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. Neuron 64, 431–439.
LeDoux, J., 2012. Rethinking the Emotional Brain. Neuron 73, 653–676.
Lempel, A., Ziv, J., 1976. On the Complexity of Finite Sequences. IEEE Trans. Inform. Theory 22, 75–81.
Liu, C., Goel, P., Kaeser, P.S., 2021. Spatial and temporal scales of dopamine transmission. Nat Rev Neurosci 22, 345–358.
Ljungberg, T., Apicella, P., Schultz, W., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology 67, 145–163.
Lohani, S., Martig, A.K., Deisseroth, K., Witten, I.B., Moghaddam, B., 2019. Dopamine Modulation of Prefrontal Cortex Activity Is Manifold and Operates at Multiple Temporal and Spatial Scales. Cell Reports 27, 99-114.e6.
Mameli-Engvall, M., Evrard, A., Pons, S., Maskos, U., Svensson, T.H., Changeux, J.-P., Faure, P., 2006. Hierarchical Control of Dopamine Neuron-Firing Patterns by Nicotinic Receptors. Neuron 50, 911–921.
Margolis, E.B., Lock, H., Hjelmstad, G.O., Fields, H.L., 2006. The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons?: Electrophysiological properties of VTA neurons. The Journal of Physiology 577, 907–924.
Maroteaux, M., Mameli, M., 2012. Cocaine Evokes Projection-Specific Synaptic Plasticity of Lateral Habenula Neurons. Journal of Neuroscience 32, 12641–12646.
Matsumoto, M., Hikosaka, O., 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459, 837–841.
Maubourguet, N., Lesne, A., Changeux, J.-P., Maskos, U., Faure, P., 2008. Behavioral Sequence Analysis Reveals a Novel Role for ß2* Nicotinic Receptors in Exploration. PLOS Computational Biology 4, e1000229.
Meiser, J., Weindl, D., Hiller, K., 2013. Complexity of dopamine metabolism. Cell Commun Signal 11, 34.
Miller, E.K., Cohen, J.D., 2001. An Integrative Theory of Prefrontal Cortex Function. Annu. Rev. Neurosci. 24, 167–202.
Miller, E.K., Wallis, J.D., 2013. The Prefrontal Cortex and Executive Brain Functions. In: Fundamental Neuroscience. Elsevier, pp. 1069–1089.
Miller, K.J., Ludvig, E.A., Pezzulo, G., Shenhav, A., 2018. Realigning Models of Habitual and Goal-Directed Decision-Making. In: Goal-Directed Decision Making. Elsevier, pp. 407–428.
Miller, K.J., Shenhav, A., Ludvig, E.A., 2019. Habits without Values. Psychol Rev 126, 292–311.
Moaddab, M., McDannald, M.A., 2021. Retrorubral field is a hub for diverse threat and aversive outcome signals. Current Biology 31, 2099-2110.e5.
Mohebi, A., Pettibone, J.R., Hamid, A.A., Wong, J.-M.T., Vinson, L.T., Patriarchi, T., Tian, L., Kennedy, R.T., Berke, J.D., 2019. Dissociable dopamine dynamics for learning and motivation. Nature 570, 65–70.
Montague, P.R., Hyman, S.E., Cohen, J.D., 2004. Computational roles for dopamine in behavioural control. Nature 431, 760–767.
Morales, M., Margolis, E.B., 2017. Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. Nat Rev Neurosci 18, 73–85.

Morales, M., Root, D.H., 2014. Glutamate neurons within the midbrain dopamine regions. Neuroscience, The Ventral Tegmentum and Dopamine: A New Wave of Diversity 282, 60–68.

Morel, C., Fernandez, S.P., Pantouli, F., Meye, F.J., Marti, F., Tolu, S., Parnaudeau, S., Marie, H., Tronche, F., Maskos, U., Moretti, M., Gotti, C., Han, M.-H., Bailey, A., Mameli, M., Barik, J., Faure, P., 2018. Nicotinic receptors mediate stress-nicotine detrimental interplay via dopamine cells' activity. Mol Psychiatry 23, 1597–1605.

Nair-Roberts, R.G., Chatelain-Badie, S.D., Benson, E., White-Cooper, H., Bolam, J.P., Ungless, M.A., 2008. Stereological estimates of dopaminergic, GABAergic and glutamatergic neurons in the ventral tegmental area, substantia nigra and retrorubral field in the rat. Neuroscience 152, 1024–1031.

Naudé, J., Dongelmans, M., Faure, P., 2015. Nicotinic alteration of decision-making. Neuropharmacology, The Nicotinic Acetylcholine Receptor: From Molecular Biology to Cognition 96, 244–254.

Naudé, J., Tolu, S., Dongelmans, M., Torquet, N., Valverde, S., Rodriguez, G., Pons, S., Maskos, U., Mourot, A., Marti, F., Faure, P., 2016. Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nature Neuroscience 19, 471–478.

Nguyen, C., Mondoloni, S., Le Borgne, T., Centeno, I., Come, M., Jehl, J., Solié, C., Reynolds, L.M., Durand-de Cuttoli, R., Tolu, S., Valverde, S., Didienne, S., Hannesse, B., Fiancette, J.-F., Pons, S., Maskos, U., Deroche-Gamonet, V., Dalkara, D., Hardelin, J.-P., Mourot, A., Marti, F., Faure, P., 2021. Nicotine inhibits the VTA-to-amygdala dopamine pathway to promote anxiety. Neuron S0896627321004281.

Nicola, S.M., 2010. 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. J. Neurosci. 30, 16585–16600.

Niv, Y., 2009. Reinforcement learning in the brain. Journal of Mathematical Psychology 53, 139–154.

Noonan, M.P., Kolling, N., Walton, M.E., Rushworth, M.F.S., 2012. Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement. European Journal of Neuroscience 35, 997–1010.

Olds, J., Milner, P., 1954. Positive reinforcement produced by electrical stimulation of spetal area and other regions of rat brain. Journal of Comparative and Physiological Psychology 419–427.

Olds, M.E., Olds, J., 1963. Approach-avoidance analysis of rat diencephalon. Journal of Comparative Neurology 120, 259–295.

Olson, V.G., Nestler, E.J., 2007. Topographical organization of GABAergic neurons within the ventral tegmental area of the rat. Synapse 61, 87–95.

Omelchenko, N., Bell, R., Sesack, S.R., 2009. Lateral habenula projections to dopamine and GABA neurons in the rat ventral tegmental area. European Journal of Neuroscience 30, 1239–1250.

Omelchenko, N., Sesack, S.R., 2006. Cholinergic axons in the rat ventral tegmental area synapse preferentially onto mesoaccumbens dopamine neurons. Journal of Comparative Neurology 494, 863–875.

O'Neill, M., Schultz, W., 2010. Coding of Reward Risk by Orbitofrontal Neurons Is Mostly Distinct from Coding of Reward Value. Neuron 68, 789–800.

Onge, J.R.S., Abhari, H., Floresco, S.B., 2011. Dissociable Contributions by Prefrontal D1 and D2 Receptors to Risk-Based Decision Making. J. Neurosci. 31, 8625–8633.

Oudeyer, P.-Y., Kaplan, F., 2009. What is intrinsic motivation? A typology of computational approaches. Front. Neurorobot. 1.

Padoa-Schioppa, C., 2009. Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. J. Neurosci. 29, 14004–14014.

Padoa-Schioppa, C., 2011. Neurobiology of Economic Choice: A Good-Based Model. Annu. Rev. Neurosci. 34, 333–359.
Paladini, C.A., Roeper, J., 2014. Generating bursts (and pauses) in the dopamine midbrain neurons. Neuroscience, The Ventral Tegmentum and Dopamine: A New Wave of Diversity 282, 109–121.
Panksepp, J., 2011. The basic emotional circuits of mammalian brains: Do animals have affective lives? Neuroscience & Biobehavioral Reviews 35, 1791–1804.
Park, J., Moghaddam, B., 2017. Risk of punishment influences discrete and coordinated encoding of reward-guided actions by prefrontal cortex and VTA neurons. eLife 6.
Parr, L.A., Waller, B.M., Fugate, J., 2005. Emotional communication in primates: implications for neurobiology. Current Opinion in Neurobiology, Motor sytems / Neurobiology of behaviour 15, 716–720.
Pascoli, V., Terrier, J., Hiver, A., Lüscher, C., 2015. Sufficiency of Mesolimbic Dopamine Neuron Stimulation for the Progression to Addiction. Neuron 88, 1054–1066.
Passingham, R.E., Bengtsson, S.L., Lau, H.C., 2010. Medial frontal cortex: from self-generated action to reflection on one's own performance. Trends in Cognitive Sciences 14, 16–21.
Patriarchi, T., Cho, J.R., Merten, K., Howe, M.W., Marley, A., Xiong, W.-H., Folk, R.W., Broussard, G.J., Liang, R., Jang, M.J., Zhong, H., Dombeck, D., von Zastrow, M., Nimmerjahn, A., Gradinaru, V., Williams, J.T., Tian, L., 2018. Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. Science 360, eaat4422.
Pessiglione, M., 2009. Motivation et récompense : le point de vue des neurosciences cognitives – Motivation and reward: views from cognitive neuroscience 6.
Pessiglione, M., 2014. Comment le cerveau motive le comportement : du circuit de la récompense au système des valeurs. Bulletin de l'Académie Nationale de Médecine 198, 1283–1296.
Pistillo, F., Clementi, F., Zoli, M., Gotti, C., 2015. Nicotinic, glutamatergic and dopaminergic synaptic transmission and plasticity in the mesocorticolimbic system: Focus on nicotine effects. Progress in Neurobiology 124, 1–27.
Plassmann, H., O'Doherty, J., Rangel, A., 2007. Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. J. Neurosci. 27, 9984–9988.
Puig, M.V., Antzoulatos, E.G., Miller, E.K., 2014. Prefrontal dopamine in associative learning and memory. Neuroscience 282, 217–229.
Puig, M.V., Miller, E.K., 2012. The Role of Prefrontal Dopamine D1 Receptors in the Neural Mechanisms of Associative Learning. Neuron 74, 874–886.
Quilodran, R., Rothé, M., Procyk, E., 2008. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. Neuron 57, 314–325.
Raghavachari, S., Lisman, J.E., Tully, M., Madsen, J.R., Bromfield, E.B., Kahana, M.J., 2006. Theta Oscillations in Human Cortex During a Working-Memory Task: Evidence for Local Generators. Journal of Neurophysiology 95, 1630–1638.
Rangel, A., Camerer, C., Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience 9, 545–556.
Rich, E.L., Shapiro, M., 2009. Rat Prefrontal Cortical Neurons Selectively Code Strategy Switches. J. Neurosci. 29, 7208–7219.
Robinson, M.J.F., Berridge, K.C., 2013. Instant Transformation of Learned Repulsion into Motivational "Wanting." Current Biology 23, 282–289.
Roesch, M.R., Calu, D.J., Schoenbaum, G., 2007. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience 10, 1615–1624.

Roesch, M.R., Taylor, A.R., Schoenbaum, G., 2006. Encoding of Time-Discounted Rewards in Orbitofrontal Cortex Is Independent of Value Representation. Neuron 51, 509–520.
Rudebeck, P.H., Izquierdo, A., 2021. Foraging with the frontal cortex: A cross-species evaluation of reward-guided behavior. Neuropsychopharmacol.
Rushworth, M.F., Kolling, N., Sallet, J., Mars, R.B., 2012. Valuation and decision-making in frontal cortex: one or many serial or parallel systems? Current Opinion in Neurobiology, Decision making 22, 946–955.
Rushworth, M.F.S., Behrens, T.E.J., 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11, 389–397.
Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., Behrens, T.E., 2011. Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70, 1054–1069.
Russo, S.J., Nestler, E.J., 2013. The brain reward circuitry in mood disorders. Nat Rev Neurosci 14, 609–625.
Salamone, J., Correa, M., Farrar, A., Nunes, E., Pardo, M., 2009. Dopamine, behavioral economics, and effort. Frontiers in Behavioral Neuroscience 3, 13.
Salamone, J.D., Correa, M., 2012. The Mysterious Motivational Functions of Mesolimbic Dopamine. Neuron 76, 470–485.
Schoenbaum, G., Takahashi, Y., Liu, T.-L., McDannald, M.A., 2011. Does the orbitofrontal cortex signal value? Annals of the New York Academy of Sciences 1239, 87–99.
Schultz, W., 1986. Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. Journal of Neurophysiology 56, 1439–1461.
Schultz, W., 1998. Predictive Reward Signal of Dopamine Neurons. Journal of Neurophysiology 80, 1–27.
Schultz, W., 2006. Behavioral Theories and the Neurophysiology of Reward. Annu. Rev. Psychol. 57, 87–115.
Schultz, W., 2007. Multiple Dopamine Functions at Different Time Courses. Annual Review of Neuroscience 30, 259–288.
Schultz, W., 2015. Neuronal Reward and Decision Signals: From Theories to Data. Physiological Reviews 95, 853–951.
Schultz, W., Dayan, P., Montague, P.R., 1997. A Neural Substrate of Prediction and Reward. Science, New Series 275, 1593–1599.
Seamans, J.K., Yang, C.R., 2004. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. Progress in Neurobiology 74, 1–58.
Sesack, R., Picke, M., 1994. Ultrastructural Localization of D, Receptor-like lmmunoreactivity in Midbrain Dopamine Neurons and Their Striatal Targets. The Journal of Neuroscience 19.
Sharpe, M.J., Stalnaker, T., Schuck, N.W., Killcross, S., Schoenbaum, G., Niv, Y., 2019. An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. Annu. Rev. Psychol. 70, 53–76.
Shi, W.-X., 2009. Electrophysiological Characteristics of Dopamine Neurons: A 35-Year Update. In: Giovanni, G., Di Matteo, V., Esposito, E. (Eds.), Birth, Life and Death of Dopaminergic Neurons in the Substantia Nigra. Springer Vienna, Vienna, pp. 103–119.
Shizgal, P., 1997. Neural basis of utility estimation. Current Opinion in Neurobiology 7, 198–208.
Shizgal, P., Conover, K., 1996. On the Neural Computation of Utility. Current Directions in Psychological Science 5, 37–43.
Soden, M.E., Jones, G.L., Sanford, C.A., Chung, A.S., Güler, A.D., Chavkin, C., Luján, R., Zweifel, L.S., 2013. Disruption of Dopamine Neuron Activity Pattern Regulation through Selective Expression of a Human KCNN3 Mutation. Neuron 80, 997–1009.
Spence, K.W., Lippitt, R., 1946. An experimental test of the sign-gestalt theory of trial and error learning. Journal of Experimental Psychology 36, 491–502.

St. Onge, J.R., Floresco, S.B., 2010. Prefrontal Cortical Contribution to Risk-Based Decision Making. Cereb Cortex 20, 1816–1828.
Stalnaker, T.A., Cooch, N.K., Schoenbaum, G., 2015. What the orbitofrontal cortex does not do. Nature Neuroscience 18, 620–627.
Stamatakis, A.M., Sparta, D.R., Jennings, J.H., McElligott, Z.A., Decot, H., Stuber, G.D., 2014. Amygdala and bed nucleus of the stria terminalis circuitry: Implications for addiction-related behaviors. Neuropharmacology, NIDA 40th Anniversary Issue 76, 320–328.
Starkweather, C.K., Gershman, S.J., Uchida, N., 2018. The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. Neuron 98, 616-629.e6.
Stauffer, W.R., Lak, A., Schultz, W., 2014. Dopamine Reward Prediction Error Responses Reflect Marginal Utility. Current Biology 24, 2491–2500.
Steinberg, E.E., Janak, P.H., 2013. Establishing causality for dopamine in neural function and behavior with optogenetics. Brain Research, Optogenetics and Pharmacogenetics in Neuronal Function and Dysfunction 1511, 46–64.
Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., Janak, P.H., 2013. A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16, 966–973.
Stellar, E., 1982. Brain Mechanisms in Hedonic Processes. In: Pfaff, D.W. (Ed.), The Physiological Mechanisms of Motivation. Springer New York, New York, NY, pp. 377–407.
Stuber, G.D., Hnasko, T.S., Britt, J.P., Edwards, R.H., Bonci, A., 2010. Dopaminergic Terminals in the Nucleus Accumbens But Not the Dorsal Striatum Corelease Glutamate. Journal of Neuroscience 30, 8229–8233.
Sul, J.H., Kim, H., Huh, N., Lee, D., Jung, M.W., 2010. Distinct Roles of Rodent Orbitofrontal and Medial Prefrontal Cortex in Decision Making. Neuron 66, 449–460.
Surmeier, D.J., Ding, J., Day, M., Wang, Z., Shen, W., 2007. D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends in Neurosciences 30, 228–235.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction 398.
Swanson, L.W., 1982. The projections of the ventral tegmental area and adjacent regions: A combined fluorescent retrograde tracer and immunofluorescence study in the rat. Brain Research Bulletin 9, 321–353.
Takahashi, Y.K., Roesch, M.R., Stalnaker, T.A., Haney, R.Z., Calu, D.J., Taylor, A.R., Burke, K.A., Schoenbaum, G., 2009. The Orbitofrontal Cortex and Ventral Tegmental Area Are Necessary for Learning from Unexpected Outcomes. Neuron 62, 269–280.
Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., Schoenbaum, G., 2011. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat Neurosci 14, 1590–1597.
Takahashi, Y.K., Stalnaker, T.A., Roesch, M.R., Schoenbaum, G., 2017. Effects of inference on dopaminergic prediction errors depend on orbitofrontal processing. Behavioral Neuroscience 131, 127–134.
Tang, W., Kochubey, O., Kintscher, M., Schneggenburger, R., 2020. A VTA to Basal Amygdala Dopamine Projection Contributes to Signal Salient Somatosensory Events during Fear Learning. J. Neurosci. 40, 3969–3980.
Tervo, D.G.R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., Karpova, A.Y., 2014. Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. Cell 159, 21–32.
Tobler, P.N., 2005. Adaptive Coding of Reward Value by Dopamine Neurons. Science 307, 1642–1645.
Tolman, E.C., Ritchie, B.F., Kalish', D., 1946. Studies in spatial learning. I. Orientation and the short-cut. Journal of Experimental Psychology 12.

Tolu, S., Eddine, R., Marti, F., David, V., Graupner, M., Pons, S., Baudonnat, M., Husson, M., Besson, M., Reperant, C., Zemdegs, J., Pagès, C., Hay, Y.A.H., Lambolez, B., Caboche, J., Gutkin, B., Gardier, A.M., Changeux, J.-P., Faure, P., Maskos, U., 2013. Co-activation of VTA DA and GABA neurons mediates nicotine reinforcement. Molecular Psychiatry 18, 382–393.

Tolu, S., Marti, F., Morel, C., Perrier, C., Torquet, N., Pons, S., de Beaurepaire, R., Faure, P., 2017. Nicotine enhances alcohol intake and dopaminergic responses through β2* and β4* nicotinic acetylcholine receptors. Sci Rep 7, 45116.

Torquet, N., Marti, F., Campart, C., Tolu, S., Nguyen, C., Oberto, V., Benallaoua, M., Naudé, J., Didienne, S., Debray, N., Jezequel, S., Le Gouestre, L., Hannesse, B., Mariani, J., Mourot, A., Faure, P., 2018. Social interactions impact on the dopaminergic system and drive individuality. Nat Commun 9, 3081.

Tritsch, N.X., Ding, J.B., Sabatini, B.L., 2012. Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. Nature 490, 262–266.

Tritsch, N.X., Sabatini, B.L., 2012. Dopaminergic Modulation of Synaptic Transmission in Cortex and Striatum. Neuron 76, 33–50.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., Deisseroth, K., 2009. Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning. Science 324, 1080–1084.

Verharen, J.P.H., de Jong, J.W., Roelofs, T.J.M., Huffels, C.F.M., van Zessen, R., Luijendijk, M.C.M., Hamelink, R., Willuhn, I., den Ouden, H.E.M., van der Plasse, G., Adan, R.A.H., Vanderschuren, L.J.M.J., 2018. A neuronal mechanism underlying decision-making deficits during hyperdopaminergic states. Nat Commun 9, 731.

Vlaev, I., Chater, N., Stewart, N., Brown, G.D.A., 2011. Does the brain calculate value? Trends in Cognitive Sciences 15, 546–554.

Voytek, B., Kayser, A.S., Badre, D., Fegen, D., Chang, E.F., Crone, N.E., Parvizi, J., Knight, R.T., D'Esposito, M., 2015. Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. Nat Neurosci 18, 1318–1324.

Wallis, J.D., 2012. Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat Neurosci 15, 13–19.

Walsh, J.J., Han, M.H., 2014. The heterogeneity of ventral tegmental area neurons: Projection functions in a mood-related context. Neuroscience, The Ventral Tegmentum and Dopamine: A New Wave of Diversity 282, 101–108.

Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rudebeck, P.H., Rushworth, M.F.S., 2010. Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. Neuron 65, 927–939.

Walton, M.E., Bouret, S., 2019. What Is the Relationship between Dopamine and Effort? Trends in Neurosciences 42, 13.

Wang, Y., Toyoshima, O., Kunimatsu, J., Yamada, H., Matsumoto, M., 2021. Tonic firing mode of midbrain dopamine neurons continuously tracks reward values changing moment-by-moment. eLife 10, e63166.

Wassum, K.M., Ostlund, S.B., Maidment, N.T., 2012. Phasic Mesolimbic Dopamine Signaling Precedes and Predicts Performance of a Self-Initiated Action Sequence Task. Biological Psychiatry, Altered Functional Connectivity in Schizophrenia 71, 846–854.

Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A., Uchida, N., 2012. Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons. Neuron 74, 858–873.

Wilson, Robert C, Geana, A., White, J.M., Ludvig, E.A., Cohen, J.D., 2014. Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General 143, 2074–2081.

Wilson, Robert C, Takahashi, Y.K., Schoenbaum, G., Niv, Y., 2014. Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron 81, 267–279.

Wise, R.A., 1996. Addictive Drugs and Brain Stimulation Reward. Annual Review of Neuroscience 319–340.
Wise, R.A., 2005. Forebrain substrates of reward and motivation. Journal of Comparative Neurology 493, 115–121.
Wise, S.P., 2008. Forward frontal fields: phylogeny and fundamental function. Trends in Neurosciences 31, 599–608.
Xia, Y., Driscoll, J.R., Wilbrecht, L., Margolis, E.B., Fields, H.L., Hjelmstad, G.O., 2011. Nucleus Accumbens Medium Spiny Neurons Target Non-Dopaminergic Neurons in the Ventral Tegmental Area. Journal of Neuroscience 31, 7811–7816.
Yamaguchi, T., Wang, H.-L., Li, X., Ng, T.H., Morales, M., 2011. Mesocorticolimbic Glutamatergic Pathway. Journal of Neuroscience 31, 8476–8490.
Yang, H., de Jong, J.W., Tak, Y., Peck, J., Bateup, H.S., Lammel, S., 2018. Nucleus Accumbens Subnuclei Regulate Motivated Behavior via Direct Inhibition and Disinhibition of VTA Dopamine Subpopulations. Neuron 97, 434-449.e4.
Yetnikoff, L., Lavezzi, H.N., Reichard, R.A., Zahm, D.S., 2014. An update on the connections of the ventral mesencephalic dopaminergic complex. Neuroscience, The Ventral Tegmentum and Dopamine: A New Wave of Diversity 282, 23–48.
Young, P.T., 1966. Hedonic organization and regulation of behavior. Psychological Review 73, 59–86.
Yu, A.J., Dayan, P., 2005. Uncertainty, Neuromodulation, and Attention. Neuron 46, 681–692.
Yun, M., Kawai, T., Nejime, M., Yamada, H., Matsumoto, M., 2020. Signal dynamics of midbrain dopamine neurons during economic decision-making in monkeys. Sci. Adv. 6, eaba4962.
Zhang, S., Qi, J., Li, X., Wang, H.-L., Britt, J.P., Hoffman, A.F., Bonci, A., Lupica, C.R., Morales, M., 2015. Dopaminergic and glutamatergic microdomains in a subset of rodent mesoaccumbens axons. Nat Neurosci 18, 386–392.

Résumé

Les comportements de prise de décision dirigés vers une récompense reposent majoritairement sur la voie méso-cortico-limbique composée de l'Aire Tegmentale Ventrale (VTA), du Noyau Accumbens et des cortex préfrontaux (OFC et PFC), également appelée « circuit de la récompense ». Ce circuit a été impliqué dans *i)* la représentation des contingences de l'environnement et la prédiction des conséquences d'un stimulus ou d'une action, *ii)* l'encodage de valeur, *iii)* l'initiation et la planification de comportements dirigés vers un but, et *iv)* la mise en place de stratégies cognitives adaptatives et flexibles. Ce système optimiserait en particulier les choix futurs vers la répétition d'actions associées aux options de plus haute valeur, ce que l'on définit comme l'exploitation. Cependant, les individus s'engagent spontanément et fréquemment dans des actions qui ne mènent pas forcément à des récompenses ou qui semblent sous-optimales, en particulier lorsque l'environnement est incertain, complexe ou changeant. On parle alors d'exploration. Les comportements de choix peuvent découler de stimuli externes ou de représentations internes à l'animal. Ces représentations peuvent être celles de la récompense potentielle ou de l'information à gagner en lien avec une option. Par ailleurs, l'exploration peut être dirigée vers une option informative ou dériver d'une stratégie de décision plus ou moins aléatoire. Il est difficile d'étudier expérimentalement les choix auto-générés pour en comprendre les bases neuronales, en particulier lorsqu'ils sont variables et exploratoires.

Au cours de ma thèse, j'ai travaillé sur une tâche dans laquelle les souris réalisent des séquences de choix binaires successifs en se basant sur une estimation de contingences afin d'obtenir des récompenses. Dans cette tâche, les choix sont auto-générés puisque les animaux déterminent eux-mêmes l'initiation, la direction, la vigueur et la structure temporelle de leurs actions. En couplant des enregistrements électrophysiologiques et des expériences de perturbation, nous avons identifié une séquence d'activités oscillatoires et de décharge des neurones dopaminergiques de la VTA, de l'OFC et du PFC qui co-encode et co-détermine à la fois le but, l'initiation et le rythme d'exécution de l'action ainsi que la stratégie de choix.

Par ailleurs, l'utilisation d'une règle récompensant les souris lorsqu'elles génèrent des séquences variables, nous a permis de découvrir que les souris sont capables d'accroitre stratégiquement la variabilité dans leurs séquences de choix en renforçant une stratégie aléatoire, sans se baser sur la mémoire des choix précédents.