Neural and cognitive bases of confirmation bias-induced interference in declarative memory performance

Christopher Stevens

université de BORDEAUX

Thèse présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE BORDEAUX

École Doctorale des Sciences de la Vie et de la Santé

Spécialité – Neurosciences

Par Christopher STEVENS

NEURAL AND COGNITIVE BASES OF CONFIRMATION BIAS-INDUCED INTERFERENCE IN DECLARATIVE MEMORY PERFORMANCE

(Bases neurales et cognitives de l'altération de la mémoire déclarative par un biais de confirmation)

Sous la direction d'Aline MARIGHETTO

Soutenue le 25 mars 2022

Membres du Jury :

Mme. RAVEL Nadine, D.R.	Université de Lyon, France	Présidente
M. COUTUREAU Etienne, D.R.	Université de Bordeaux, France	Examinateur
M. MALLERET Gaël, C.R.	Université de Lyon, France	Rapporteur
M. PALMINTERI Stefano, C.R.	École Normale Supérieure, Paris, France	Rapporteur
Mme. MARIGHETTO Aline, D.R.	Université de Bordeaux, France	Directrice de thèse
M. MARSICANO Giovanni, D.R.	Université de Bordeaux, France	Invité

Bases neurales et cognitives de l'altération de la mémoire déclarative par un biais de confirmation.

Le biais de confirmation consiste en un phénomène cognitif bien caractérisé et universel par lequel de nouvelles informations provenant de l'environnement sont surévaluées quand elles confirment et sous-évaluées quand elles infirment un contenu cognitif préalablement consolidé (p.ex., croyances, règles d'association). Les réponses mal-adaptées que ce phénomène peut générer font partie de problèmes sociaux tels la rediffusion des «fake news» et varient selon la complexité du contexte et l'état mental du sujet.

Malgré ces faits, il existe peu d'études dédiées à l'exploration des mécanismes neuraux ou à l'évolution de ce phénomène. Ainsi, nous avons conçu un modèle murin des comportements de type «biais de confirmation» afin de nous permettre d'explorer ses substrats cognitifs et neuronaux et leur évolution. Nous avons basé notre modèle sur une définition cognitive du phénomène; surévaluation de nouveaux éléments environnementaux quand ils confirment, et sous-évaluation quand ils infirment, un contenu cognitif préalablement consolidé. Nos résultats jusque-là (employant un protocole à deux tâches et à deux contextes sur le labyrinthe radiaire) révèlent un fort effet de biais qui se manifeste comme une altération dans la performance d'une tâche de mémoire déclarative et dont la persistance varie en fonction de la complexité de l'essai.

Grâce à des analyses comportementales détaillées, nous avons su identifier des composants cognitifs plus basiques qui impactent cet effet de biais, tels l'oubli adaptatif ou l'équilibre exploration/exploitation. Ces composants sont fortement liés avec des circuits neuronaux spécifiques dont l'activité est susceptible d'être enregistrée ou modifiée in vivo. Ils sont aussi impliqués dans plusieurs troubles mentaux (dépression, schizophrénie), faisant du modèle un nouvel outil pour la recherche pré-clinique, dont nous développons des versions humaine (pour recherche clinique), et computationnel (pour formuler des prédictions à tester).

Neural and cognitive bases of confirmation bias-induced interference in declarative memory performance.

Confirmation bias is a well-described and ubiquitous cognitive behavior whereby novel information from the environment is over-valued when it confirms and under-valued when it disconfirms previously consolidated cognitive content (e.g. beliefs, learned associations, etc.). The maladaptive responses this phenomenon can give rise to are implicated in social problems such as the spread of "fake news" and vary according to both contextual complexity and the mental state of the subject.

Nevertheless, very little research has been dedicated to understanding the neural mechanisms or evolution underpinning this spontaneous human cognitive response to novel information. Thus, we designed a mouse model for confirmation bias-like behavior, enabling exploration of its cognitive and neurobiological underpinnings and their evolution. Our model is based on a cognitive level definition of the phenomenon; over-valuation of novel environmental elements which confirm and under-valuation of novel environmental elements which disconfirm a previously consolidated cognitive content. Our results to this point (using a two-task, two-context radial maze protocol) show a strong bias effect which is observable as a deviation in the performance of a classical declarative memory task, the persistence of which is trial-complexity dependent.

Detailed behavioral analysis has enabled us to identify several more basic cognitive components impacting the bias effect, such as adaptive forgetting and the exploration/exploitation balance. These cognitive components have been identified with specific neural circuits whose activity is susceptible to intervention and/or monitoring in freely moving task-performing animals. They are also implicated in many psychiatric conditions (depression, schizophrenia, etc.) making of this model a novel tool for pre-clinical research of which we are developing a human version for clinical research and a computational version for formulating and testing predictions.

Unité de Recherche

Université de Bordeaux
Neurocentre Magendie – Inserm U1215
Equipe « Physiopathologie de la Mémoire Déclarative »
146 Rue Léo Saignat
33077 Bordeaux
France

Financement

Ce travail a été soutenu par :

- Une bourse du Ministère de l'Enseignement Supérieur et de la Recherche de 2017 à 2020
- Une bourse de fin de thèse de « LabEx BRAIN »

Remerciements

Cela fait déjà six ans et demi depuis que j'ai atterri, tel un ovni venu du monde de la philo, à l'Université de Bordeaux en M1 de neurosciences. Six ans depuis la première fois que j'ai mis les pieds au Neurocentre Magendie pour mon premier stage… puis mon deuxième stage, puis ma thèse. On croise beaucoup de monde en six ans et demi et quand on ne sait jamais exactement ce qu'on est en train de faire, on finit par demander beaucoup d'aide. Résultat ; trop de monde à remercier, mais je vais essayer.

Avant tout, je tiens à remercier tous les membres du jury de ma thèse ; les docteurs Stefano Palminteri et Gaël Malleret pour leur lecture du manuscrit, ainsi que les docteurs Nadine Ravel et Etienne Coutureau pour leurs retours en tant qu'experts dans leurs domaines.

Ensuite je dois remercier tout d'abord mes guides scientifiques et intellectuels, dont j'ai appris et continue à apprendre tant de choses : Aline Marighetto, qui m'a ouvert les portes de son labo et m'a fait confiance sans trop savoir ce qui allait en découler ; Giovanni, qui m'a toujours fait sentir chez moi dans la grande famille Marsicano (aidé en cela, comme avec tout, par Astrid bien sûr !) ; Barbara Stiegler, ma directrice du côté de la philosophie, qui était le premier à me mettre en tête l'idée de poursuivre mon objectif d'étudier les neurosciences ici à Bordeaux, et donc grâce à laquelle j'ai pu vivre ce chapitre – qui m'est encore à peine croyable – de ma vie.

Il faut aussi ajouter toutes les autres personnes qui se sont mobilisées afin de m'aider à me reconvertir en neuroscientifique à partir de mon éducation philosophique : Marion Worms, Jacques Dubucs, Cédric Brun, Jacques Micheau et Daniel Voisin.

Que ce soit en cours, en stage, ou en discussion autour d'un café, il y a certains individus qui m'ont transmis une quantité incroyable de connaissances, simplement de par leur enthousiasme pour le savoir scientifique. Du côté de mes profs ; Elena Avignone, Aude Panatier, Agnes Nadjar, Daniel Voisin, Philippe De Deurwaerdere ; pendant mes deux stages au Neurocentre Magendie ; Cyril Herry, Cyril Dejean, Danny Jercog, Stéphane Valero, Ana SantAna, Bastien Redon et surtout, surtout Francis Chaouloff.

Il y a aussi ceux (en l'occurrence *celles*) qui m'ont communiqué, toute en douceur, comment se sentir à l'aise avec les souris et comment les traiter avec le respect qu'elles méritent ; Hélène Wurtz, Nânci Winke et Eva Ducourneau.

Tout le monde, surtout de la famille Marsicano, qui m'ont si bien aidé à préparer le concours de l'école doctorale (et à bien d'autres moments aussi !) : Arnau, Edgar, Roman, Antonio, Geoffrey, Luigi, José, Francis, Bastien, Mari-Carmen.

A ce point il faudrait aussi dire un très, très grand merci à tous les gens qui ont fait la grandeur de l'équipe Marsicano pendant le temps que j'ai pu passer avec, pendant mon stage de M2 avec Francis et puis tout au long de ma thèse, entreprise en collaboration avec l'équipe. Vous êtes vraiment trop nombreux pour vous nommer tous, mais, hormis ceux déjà mentionné, tout particulièrement Yamuna, Marjorie, Su, Christina, Imane,

Ignacio, Virginie, Ula et Emma qui m'ont à un moment donné été d'un aide ou d'un soutien précieux.

Tous les membres de l'équipe Marighetto où j'ai eu la chance de mener ce projet ambitieux. Là, en revanche, on n'est pas beaucoup, mais on compense bien ! Après Aline, je voudrais surtout remercier Eva et Azza sans lesquelles j'aurais été perdu tellement de fois. Elles ont toujours su supporter mes demandes d'aide avec le sourire, que j'aime croire était sincère ! Mais à un moment ou un autre pendant ma thèse toutes les membres de l'équipe ont su m'aider ou m'encourager avec ma recherche : Aline Desmedt, Nicole, Mylène, Valérie, Nathalie, Shaam et Sophie. C'était un vrai plaisir de passer ce 4 ans avec vous.

Pendant ma thèse, j'ai eu l'opportunité de surveiller quatre stagiaires dont le sérieux et l'enthousiasme ont propulsé ma recherche bien avant : Mathilde Bouchet, dont le stage si bien démarré a malheureusement été interrompu par la crise sanitaire ; Cathy Lacroix, avec laquelle nous avons fait des progrès incroyables pour rattraper le temps perdu par cette crise ; et puis Faustine et Blandine qui sont venues en stage d'observation pendant leur licence de biologie, toutes les deux pleines de questions perspicaces dont la réflexion pour y répondre a souvent ouvert le chemin vers de nouvelles découvertes.

Pendant tout ce temps que j'ai passé à Magendie, je n'oublierai certainement pas de remercier toute l'administration et toute l'équipe qui fait tourner la recherche que nous y menons. Tout d'abord, tous les animaliers, ainsi que Nathalie, Sara, Julie et Cédric, qui veillent sur le bien-être des animaux sur les épaules desquels notre recherche est montée. Ensuite, tout le monde du côté de l'administration et particulièrement, en ce qui me concerne, Poun, Franck, Dania et Stéphane Oliet, un directeur d'institut qui prend toujours le temps pour échanger et pour écouter.

Je voudrais aussi remercier les membres de l'équipe pédagogique de l'Université de Bordeaux qui m'ont confié des missions pédagogiques, à savoir Karine et Muriel, ainsi que Pascal Fossat.

Il faut une mention spéciale pour Thomas Pradeu, désigné mon tuteur de thèse mais auquel je serai à jamais reconnaissant par ailleurs puisque c'était en assistant à ses cours de philosophie de biologie en master de philosophie des sciences à la Sorbonne que s'est semée dans mon esprit la première graine d'une idée d'orienter ma recherche vers la (neuro)biologie.

"Now, if you don't mind, I'll continue in English."

Huge thanks to those, near and far, who contributed to discussion, inspired perspectives, or began to reflect on collaborations: Nicolas Rougier, Frederic Alexandre, Snigdha Dagar, and the whole Mnemosyne team; Guillaume Ferreira, Shauna Parkes, Mael Lemoïne, Tomonori Takeuchi for in-person exchanges; Cormac Doherty and Yohan John for discussions over zoom; and for stimulating and educational email exchanges, Lisa Genzel, Michael J. Frank, Michael Anderson, Allen Neuringer.

Every member of the Stackoverflow community who responded to one of my dozens of coding queries, thereby enabling me to hold onto my hair for another couple of years at least.

Similarly, the entire open science community who have been a constant inspiration and beacon. Again, too many to mention, but in particular Bernard Rentier, Björn Brembs, and especially the still much missed Jon Tennant, who wrapped the entire open science message up in, "It's just science done right." May we in turn do right by your memory.

My mum and two sisters, Siobhán and Aislinn, for having supported and encouraged me all throughout my decision to return to university, not to mention the many unexpected turns that followed on from that first fatal decision… such as somehow ending up doing a PhD in neuroscience! And, forever and always, endless gratitude to my uncle James 'Paddy' Stevens who first attuned my eyes and mind to the world as a place of boundless mystery and wonder.

All the friends and colleagues in and around science who have kept me sane over the last six and a half years by entertaining the little bursts of insanity I needed to release from time to time to do that: Eva, Oscar, Maxime, Tifenn, Oriana, Ha-Rang, Vernon, Alex, Yohan, Antonio, Ilaria, Suzanne, Weronika, Louisa, Tomas, Alessandro, Ashley, Clément.

And finally, and most of all, to my wonderful partner Mari-Carmen and our incredible son Tristan-Alejandro, thank you is far too narrow a concept. Never has such an unmitigated miracle arrived at a worse time than in the middle of a PhD and just before the worst pandemic the world has known in a century. Yet somehow we managed to emerge from it all closer than ever before, which is the greatest result anyone, scientist or otherwise, could hope for.

Publications

Etudes publiées dans des revues à comité de lecture :

Medrano M-C, Hurel I, Mesguich E, Redon B, **Stevens C**, Georges F, Melis M, Marsicano G, Chaouloff F.

Exercise craving potentiates excitatory inputs to ventral tegmental area dopaminergic neurons. **Addiction Biology**, 2020.

Muguruza C, Redon B, Fois GR, Hurel I, Scocard A, Nguyen C, **Stevens C**, Soria-Gomez E, Varilh M, Cannich A, Daniault J, Busquets-Garcia A, Pelliccia T, Caillé S, Georges F, Marsicano G, Chaouloff F.

The motivation for exercise over palatable food is dictated by cannabinoid type-1 receptors. **JCI Insight**, American Society for Clinical Investigation, 2019, 4 (5).

Etudes en préparation pour publication dans des revues à comité de lecture :

Christopher Stevens, Cathy Lacroix, Yamuna Mariani, Giovanni Marsicano, Aline Marighetto.

Indoctrination as active inhibition of spontaneous exploration: Introduction of a novel mouse model. (En préparation)

Christopher Stevens, Cathy Lacroix, Mathilde Bouchet, Faustine Roudier, Yamuna Mariani, Giovanni Marsicano, Aline Marighetto.

Investigating hallmarks of 'myside' confirmation bias in a novel mouse model of everyday-like rule revision. (En préparation)

Communications Scientifiques

Janvier, 2021	Intervenant au séminaire doctoral *Pandémie, Sciences et Société* de l'école doctorale Montaigne Humanités. Titre de l'intervention : « Epistemic pandemic – A Clash of Crises ».
Novembre, 2020	Intervenant au séminaire *Démocratie, science et éducation*, Université Bordeaux-Montaigne. Titre de l'intervention : « Démocratie et neuropédagogie ».
Mai, 2019	Présentation de poster à *NeuroFrance 2019*, Marseille. Titre : « A mouse model for exploring the neurobiology of confirmation bias-like behaviour ».
Octobre, 2018	Modérateur de séance et intervenant à *Donders Discussions 2018*, Donders Institute, Nijmegen, Les Pays-Bas. Titre de la séance modérée : « Predictive brains and biased eyes ». Titre de l'intervention : « A mouse model for exploring the neurobiology of confirmation bias-like behaviour ».

Encadrement d'Etudiants

2021 — Encadrement d'un stage de M1 en neurosciences (5 mois).

Projet : « Exploration du rôle des récepteurs CB1 sur les voies directe et indirecte des ganglions de base dans la révision des croyances ».

2020 — Encadrement d'un stage de M2 en neurosciences (5 mois).

Projet : « Exploration du rôle des récepteurs CB1 hippocampiques dans la mémoire déclarative ».

Responsabilités liées à la vie du campus pendant la thèse

2018 – 2021 — Chargé de cours au sein de l'UFR de Biologie à l'Université de Bordeaux

64 heures de TD assurées, dont :

L3 biologie, « Cognition et émotion dans le royaume animal »,

(Responsable – Karine Massé), 2018-2021.

M1 neurosciences, « Démarche expérimentale en neurosciences »,

(Responsable – Muriel Darnaudery), 2019.

2018 – 2021 — Représentant des étudiants (masters, doctorants et post-docs) du Neurocentre Magendie au conseil du département Bordeaux Neurocampus.

2018 – 2021 — Référent bien-être animal de l'équipe Marighetto.

2018 – 2021 — Représentant élu des doctorants du Neurocentre Magendie au conseil de l'institut.

TABLE OF CONTENTS:

Présentation en français de l'objectif de ce projet

Afin de tirer son profit maximal d'un monde qui se situe sur un continu allant de la stabilité et de la prédictibilité d'un côté jusqu'à l'imprévisibilité de la nouveauté et du changement de l'autre, l'organisme doit arriver au meilleur équilibre entre l'exploitation de ce qui lui paraît stable et prédictible et l'exploration de ce qu'il perçoit comme incertain ou nouveau. Ainsi, à la rencontre du nouveau, l'organisme doit choisir s'il va s'y accommoder, ajustant son état interne afin d'appréhender la nouveauté en tant que telle, ou bien s'il va plutôt tenter de l'assimiler, de le subordonner à ses pré-acquis, à son état interne préétabli. Chez l'humain, ce phénomène se trouve non seulement au niveau de la perception mais aussi au niveau du raisonnement. Dans ce dernier cas, quand on cherche à assimiler du nouveau à de l'ancien, on parle d'un « biais de confirmation » car il s'agit de favoriser ces éléments du nouveau qui confirment nos acquis tout en dévalorisant ceux qui les infirment. Aujourd'hui, on en parle beaucoup dans le contexte de la diffusion des « fake news », de la crise de la reproductibilité, etc. En effet, plus nous sommes confrontés à des flux d'information, plus le biais de confirmation devient un enjeu pour la société. Il est donc de plus en plus important de comprendre les bases neurobiologiques de ce phénomène, lesquelles sauraient indiquer les meilleures méthodes cognitives pour le surmonter, surtout chez les populations âgées ou atteintes de dépression, etc., qui en sont plus susceptibles.

L'utilisation de modèles animaux dans l'exploration des substrats neuraux de diverses conditions physiopathologiques est l'une des clefs de voûte des neurosciences, or jusqu'ici il n'existe pas de modèle animal du biais de confirmation. Ainsi, afin de répondre à ce manque, dans ce projet nous nous sommes données l'objectif de profiter des tendances avérées de la souris, d'une part, à l'exploration spontanée et, d'autre part, à l'exploitation acquise. A partir de ces bases nous avons conçu et validé un modèle comportemental qui fait émerger chez la souris une phénotype comparable, à bien des égards, au biais de confirmation. Ce modèle, nous avons pu par la suite l'utiliser pour entamer une investigation des bases neurobiologiques et évolutionnaires de ce biais si présent dans le monde d'aujourd'hui. Ce sont les résultats présentés dans ce projet de thèse.

Preface

The desire and motivation to dedicate my thesis in neuroscience to the study of confirmation bias relates directly to a broader research question which has been the focus of my work since the first year of my master's degree in philosophy of science and the mind at La Sorbonne, Université Paris IV. It is the question of scientific education itself, specifically as this relates to the transition from one theoretical perspective of a given phenomenon to another: rule revision up to the most abstract levels. The subject of my master 1 mémoire was the philosophy of cognitive dissonance, a phenomenon the proximity of which to 'myside' confirmation bias is evoked within the present manuscript. It was while researching my master 2 mémoire on the subject of pluralism in scientific education, however, that my focus first began to be drawn towards neuroscience and the evolution of the higher cognitive functions. The work I present here constitutes a giant leap forward for my research in this direction, as well as a small step towards a shared deeper understanding of the evolution and neurobiology of our modes of reasoning and educating. It remains for me, in the future, to reflect upon and draw out the philosophical implications of the results these last four years of PhD research have given me the opportunity to produce. But for the present moment, it is still to its scientific implications my attention is turned.

The findings of my research are presented here in the form of two "expanded" articles currently in preparation for publication. The first of these is an investigation of how mice respond to an indoctrination-like protocol of learning, i.e. one which explicitly discourages their spontaneous drive to explore, their "curiosity," and which is specifically the type of learning susceptible, in humans, to later give rise to 'myside' confirmation bias. The second then investigates the confirmation bias-like behavior effectively elicited in mice who have learnt a rule in such an indoctrination-like manner when they are subsequently brought into a novel environment where this rule must be revised.

General Introduction

> *"Confirmation bias has been used in the psychological literature to refer to a variety of phenomena. Here I take the term to represent a generic concept that subsumes several more specific ideas that connote the inappropriate bolstering of hypotheses or beliefs whose truth is in question."*
>
> Raymond Nickerson, (1998).

The empirical findings from the two studies constituting the present PhD research will ultimately inspire the interpretation that the object of investigation, confirmation bias, can be meaningfully theorized as a particular product or artefact of organisms possessing multiple memory and learning systems having to navigate dynamic environments that demand revision of previously formed state-action policies. It seems judicious, therefore, to open proceedings with an introduction briefly outlining the history, development, and relevance to the present research endeavor of the central technical terms: 1) state-action policies; 2) multiple memory and learning systems, and; 3) confirmation bias itself. My hope is that, over the course of this introduction, it will become clear to the reader that the now uncontroversially admitted presence of 1) and 2) in a vast range of species naturally gives rise to two key questions regarding 3), i.e. confirmation bias, being a phenomenon which, by contrast, the literature has thus far admitted of only in humans. Those two questions are:

1. Do non-human animals whose state-action policies are shaped via multiple memory and learning systems also, putatively *thereby*, possess the cognitive capacity to manifest confirmation bias-like behaviors?

2. Is the well-characterized phenomenon of confirmation bias in humans a consequence of *our* state-action policies being shaped via multiple memory and learning systems?

The first question neatly sums up the orientation of the research I have undertaken during my PhD, and my findings in this respect, communicated in the two research articles constituting Part 1 and Part 2 of this manuscript, represent the first elements of an empirical response to it to appear in the literature. In turn, I hope my present contributions will inspire future research to tackle the second question in a similarly direct and empirical manner.

1. State-action policies

Since I do not actually develop a computational approach in the present work, my borrowing of and reflections around certain terms and notions from the domain of computational reinforcement learning is primarily intended as an aid in conceptualizing the extent to which certain complex cognitive functions displayed by both humans and non-human animals are eminently comparable and mutually informative. This is not, however, a purely neutral consideration, since certain philosophical positions, either implicitly or explicitly but in either case *widely* held, make many skeptical or dismissive of the idea that 'beliefs' are something animals are capable of possessing. This presents an unignorable obstacle in the context of presenting an animal model of any human cognitive process, such as confirmation bias, which is inextricably intertwined with beliefs.

By providing conceptual language and tools for grouping together all cognitive content that directs action in a context-dependent manner, reinforcement learning isolates and unifies what beliefs, rules, strategies, stimulus-response behaviors, memory- and learning-

based decisions, reward-expectation based probabilistic choices, and more all have in common: prior learning recalled, via perception of current environmental state, as a guide to action. Indeed, computational reinforcement learning theories have already been applied to make it easier for us to isolate and analyze the *general* cognitive conditions underpinning phenomena otherwise specifically associated with humans, such as indoctrination and confirmation bias notably (Palminteri, 2021; Palminteri et al., 2017; Summerfield & Parpart, 2022). This can in turn facilitate the work involved in designing animal models capable of eliciting behaviors which, if observed, would thereby imply the presence of comparable cognitive conditions in the species in question.

The following presentation of the concept of state-action policies does not aim to be exhaustive, nor even comprehensive, as to do this would require delving deep into domains such as dynamic programming, which are both beyond the expertise of the author and graciously not necessary for the reader to grasp in order to fully understand the experimental approach adopted here. Rather, my intention is simply to give the reader a sense of how mutually beneficial familiarity with concepts from both experimental psychology *and* reinforcement learning can be. For similar reasons, I have made the choice to exclude mathematical annotation from this brief introduction, as I have learnt from personal experience that it can present a seemingly insuperable psychological obstacle to the uninitiated.

1.1 An open-ended history.

At its most basic, a state-action policy is a formalism from the language of reinforcement learning that describes any kind of decision-making rule or strategy consisting in "a mapping from perceived states of the environment to actions to be taken when in those states" (Sutton & Barto, 2014). How reinforcement learning conceptualizes this mapping is historically rooted in experimental animal psychology, notably in the works of Edward

Thorndike. Thorndike (1911) famously presents the simple but powerful concept of the Law of Effect, the idea that any action leading to an outcome the agent perceives as positive will increase the probability of the agent repeating that action, whereas any action leading to an outcome the agent perceives as negative will decrease the probability of that action being repeated. These are what are now commonly referred to as, respectively, *positive reinforcers* and *negative reinforcers*, terminology made famous in the early work of the behaviorist experimentalist and theorist B.F. Skinner (B F Skinner, 1938)[1].

Since the units reinforcement modulates are initially spontaneous actions, the Law of Effect relates to what is called *instrumental* learning[2]. This in turn implies an innate, or primitive, trial-and-error strategy on the part of the agent with respect to its environment: execute an action; evaluate its outcome; increase or decrease frequency of action as per outcome evaluation. As an illustration, in an experimental environment the action might be pressing a lever (initially as an action produced at random), the outcome a food reward evaluated as positive, and the consequence of this positive reinforcement an increased comparative probability of pressing the lever again rather than engaging in some other non- or negatively reinforced action. What this implies is a learning mechanism that relies equally on 1) exploration (or *searching*, i.e. trying out various actions, or more accurately

[1] I intentionally side-step the debate between common and technical usages of the terms "positive reinforcer" and "negative reinforcer." Skinner himself initially used these terms to refer to reward and punishment, respectively, only later aligning himself with a more technical use that is now standard in the psychology literature. There, the label "negative reinforcer" is not applied to punishment but rather to the absence ("negative") of an aversive stimuli, such as pressing a lever to avoid mild electric shock. In other words, in the technical usage, "positive" and "negative" are not to be understood as "good" and "bad" but rather analogously to how the same terms are used when speaking of, for example, symptoms of mental disorders: "positive" symptoms are an addition, such as delusions; "negative" symptoms are a subtraction, such as social withdrawal. Within the domain of computational reinforcement learning, one generally finds only the earlier and more everyday usage applied: positive reinforcer = higher reward value; negative reinforcer = lower or minus reward value. This makes sense since computational reinforcement learning reduces all notion of reward and punishment to relative numerical scalar values.

[2] Instrumental learning stands in opposition to classical Pavlovian learning, whereby delivery of a reinforcer (food, punishment, etc.) is cued, independently of the animal's own actions, by an initially neutral stimulus (bell, tone, light, etc.). Over time, this stimulus becomes cognitively associatively paired with the actual reinforcer such that stimulus presentation will begin to provoke a similar neurophysiological reaction as reinforcer delivery itself: the bell makes the dog salivate, the tone makes the mouse freeze, etc.

*inter*actions with the environment), 2) evaluation (of action outcomes), and 3) associative memory (i.e. storing, for future recall, previous action-outcome-evaluation associations). Since these three components readily lend themselves to geometric and numeric abstraction, it is easy to understand why Thorndike's theorization of trial-and-error learning went on to inspire the still nascent discipline of artificial intelligence (AI) in the 1950s.

By various accidents of history, AI research became decoupled from and progressed during several decades without further consideration of animal psychology or cognition (Gershman et al., 2015). Underlining this separation in his groundbreaking advancement towards re-bridging that gap, Chris Watkins commented in his PhD thesis in 1989 that he did not know of "a single paper on animal learning published in the main stream of literature on 'artificial intelligence'" (Watkins, 1989). The particular sensitivity to matters of ecological learning this observation reveals has as a result that Watkins' work (Watkins, 1989; Watkins & Dayan, 1992) is particularly interesting for those whose background is in the domain of animal research rather than computation or AI. This is because Watkins takes as his starting point the conviction that deep reflection on how animals learn to behave efficiently in real environments (ecological or experimental) could (and indeed *did*) inspire great progress in the domain of computational reinforcement learning. In turn, in the domain of neuroscience where the behavioral dimension is regularly accused of being neglected (Krakauer et al., 2017; Niv, 2021), recent successes of reinforcement learning might inspire us with respect to the potential returns of deeper reflection on the behaviors of our own preferred animal models. As an example, in his conceptualization of the problem of reinforcement learning, Watkins succeeds in cutting through debate over the nature of the complex, putative relationships between instrumental and classical Pavlovian learning mechanisms elicited by highly constrained experimental environments by instead reframing the question in evolutionary terms, asking; by what *general* learning mechanisms

might an animal in a given environment modify its behavior in accordance with the optimization of its present and future reproductive success? As he promptly points out, however, the question of how to identify and define what is *optimal* for a given animal agent is no simple affair, especially when dealing with agents who have evolved in naturalistic *dynamic* environments and when furthermore lacking knowledge about potentially relevant innate and context-dependent behavioral tendencies resulting from that evolution (Summerfield & Parpart, 2022; Watkins, 1989). This in turn adds a layer of complexity when it comes to evaluating, in an observational capacity, whether or not, and at what scale of reference (immediate task? lifetime? evolutionary?), a given state-action policy can be said to be 'optimal'. Indeed, this stands as an important open question for investigation at the crossroads of AI, cognitive science, and neuroscience, one which will be further discussed in the course of the present manuscript.

1.2 Policies: learned, innate, revised.

Through reinforcement learning, we gain formalisms for accounting not only for how policies can be formed but also for how they can be revised. In both cases, this is understood to be the result of the agent evaluating outcomes (which may be fixed or dynamic) from actions taken when in a given state, associatively storing these state-action-outcome evaluations, and using them to inform future action when in the same or a similar state. As such, the term state-action policy allows us to subsume, under one abstract concept, any plastic (i.e. revisable) cognitive content that is understood to govern an organism's (i.e. agent's) action selection in a context-dependent manner. Cognitive content such as beliefs, rules, attitudes, stimulus-response associations, etc.

Central to the concept of a state-action policy is the fact that each action taken also brings the agent into a new state. This has been referred to as SARSA, for *state-action-reward-state-*

action, whereby from an initial state s_1 the agent takes action a_1 whereupon it receives reward r and moves into state s_2, from where it can take action a_2, and so on (Sutton & Barto, 2014). This can be illustrated using a well-known example which, in Part 1 of the present work, will be referred to as a state-action policy that emerges spontaneously in mice under specific laboratory conditions, i.e. spatial alternation (Dember & Richman, 1989; Richman et al., 1986).

Spatial alternation has been classically studied using either T- or Y-maze apparatuses. These consist in a starting corridor leading to a choice-point, being the physical junction where a choice must be made to visit either the left or the right arm of the maze. In a free choice version of the task, an animal placed at the base of the starting corridor will first advance towards the choice-point. Let state s_1 be the first arrival of the mouse at the choice-point[3]. From this state it can choose as an action either to explore the left arm or the right arm. Let us suppose it chooses the left arm and let us call this a_1. In a reinforced version of the task, the mouse will receive usually a food reward r at the end of the left arm it has just explored. Following consumption of the reward, the animal is returned by the experimenter to the starting corridor. When it arrives again at the choice-point, this now represents a new state we can call s_2, comprised of both the animal's location at the choice-point *plus* the stored memory that its previous relevant state-action a_1 was to explore the left arm. The animal's innate spatial alternation policy dictates that the most probable state-action a_2 that the mouse will take now is to explore the previously unexplored right arm. If the experimenter is reinforcing spatial alternation, then on this trial choosing the right arm will be rewarded (positively reinforced) and choosing the left arm not rewarded (negatively reinforced), and so when moved to s_3 (location *plus* the stored memory that previous state-action a_2 was to explore the right arm) the mouse's next state-action a_3 will

[3] Note that reinforcement learning algorithms allow for an essentially limitless range in the scale of what counts as a state or action. For illustrative purposes, here we zoom out to the scale of only the most strictly necessary task-definition relevant choice actions.

most likely be to explore the left arm again, and so on. The complete state-action policy for this spatial alternation reinforcing T- or Y-maze experimental environment can thus be described something like this; when in a choice-point state s_n, take that state-action a_n which is the complement of the state-action a_{n-1} taken in state s_{n-1}, where the set A of all possible actions the agent can choose from is limited to {'explore left arm', 'explore right arm'}.

By merit of being a reliable *spontaneous* behavioral tendency, presumably preserved across evolution due to some reproductive advantage it brings to the organism, the case of spatial alternation calls for special consideration, falling under what Watkins refers to as "innate knowledge." In the context of learning and the evaluation of learning rates, the fundamental question he asks is this: "What types of innate knowledge do animals have, and in what ways does this innate knowledge contribute to learning?" (Watkins, 1989). However, it is furthermore just as important to frame such a notion of "innate knowledge" as it relates to the behavioral affordances provided by a given environment. For example, mice will spatially alternate in a T- or Y-maze even if this behavior is *not* positively reinforced, meaning this particular state-action policy emerges even in the absence of an explicit environmental reinforcer to evaluate. In fact, recent work has shown that mice will spatially alternate in a T-maze even after prior establishment of a preference for a reward found in only one of the arms (Habedank et al., 2021).

This latter observation supports a theory of animal exploration wherein global *information gain* takes primacy over foraging, in the strict sense, as the principal cognitive drive underpinning exploratory behavior (Inglis et al., 2001). This primacy of pure exploration can even be related to Jaak Panksepp's theorization of "seeking" as the most fundamental *affective* drive of organisms, "which helps elaborate energetic search and goal-directed behaviors in behalf of any of a variety of distinct goal objects" (Panksepp, 1998). Foraging specifically for food, in this theory, becomes just one special case of a global exploratory

drive, primitive with respect to any particular goal: sometimes exploration may take the form of foraging, other times mate- or shelter-seeking, etc. Through these interpretative lenses, it seems more accurate to affirm that what mice do spontaneously is not so much to spatially alternate as it is just to *explore*. In this interpretation, it is then the physical conditions, if not to say constraints, of the T- or Y-maze environments which channel this exploration to manifest as what experimenters subsequently observe and label as 'spontaneous spatial alternation'. Indeed, in terms of reinforcement learning, all other things being equal, spatial alternation can be understood simply as the maximally efficient or optimal policy for exploring a T- or Y-maze.

Conversely, it is by this same "innate knowledge" policy logic that in Part 1 of the present work, where the environmental conditions of the tactile discrimination task reinforce explicitly *non*-exploratory behavior, we will interpret this learning not as *initial* formation of a novel policy but rather as demanding a context-dependent *revision* of the innate exploratory policy. As an illustration, let us briefly elaborate how this relates to experimental conditions employed in the present investigation. In the tactile discrimination experimental set-up presented in Part 1, the surface area of the radial maze is divided according to two different surface types, one smooth, one irregular. Since the experiment is conducted in darkness, in the absence of *visual* spatial clues, an efficient strategy for ensuring exploration of the whole environment would therefore be to form the state-action policy of alternating surface type chosen when deciding, trial by trial, which to visit between two neighboring arms of the radial maze, each of which has a different surface type. In this context, we can imagine a state s_2 (location at choice-point *plus* the stored memory that previous state-action a_1 was to explore, say, a smooth surfaced arm) in which the most efficient state-action a_2 the mouse can take, if acting according to the innate exploratory state-action policy, will be to now choose the irregular surfaced arm. However, as is the actual case in our protocol, if only one of these surfaces is ever rewarded,

then exploratory behavior will be *negatively* reinforced whenever a state-action choice brings the mouse to visit an arm of the unrewarded surface. According to reinforcement learning theory, this should set in place an incremental revision, via ongoing action-outcome evaluation, of the exploratory, surface-alternation policy.

However, the crucial point to grasp here is that the behavioral manifestation of the "innate knowledge" elicited in mice by the radial maze (i.e. prioritize exploration of unexplored or least recently explored areas) does not so much contribute as it *stands in opposition* to the learning our tactile discrimination protocol aims to transmit (i.e. 'Choose only one surface'). Not to mention that, behind this opposition, is nothing less than the momentum of countless millennia of evolution. A stark contrast therefore appears with respect to behavioral tasks (such as those we present in Part 2) which are designed to *exploit* spatial alternation: here, exactly the same innate knowledge that opposes non-exploratory learning becomes essentially sufficient for successful performance. Reflection on the mutual implications, for animal behavior studies and for reinforcement learning, of this context-dependent contrast in how innate tendencies manifest poses a particularly interesting challenge to our understanding of learning and the shaping of *optimal* state-action policies on the basis of that learning, as we shall now see.

1.3 Challenges for optimality.

One of the challenges for an optimality approach to reinforcement learning, a challenge broached by Watkins and further underlined by the results from our own experiments in Parts 1 and 2 of the present work, is that what is optimally efficient in one environment may not be optimally efficient across the lifespan of an organism, who may well have to confront and overcome survival threatening changes to its environment during that time. Indeed, even though in our tactile discrimination protocol (Part 1) we extensively and unambiguously discourage mice from exploring, in what we call an "indoctrination-like"

manner, we nevertheless observe robust evidence that the exploratory drive does not so much diminish over the course of this training as it becomes progressively *actively inhibited*. In this interpretation, it is increased engagement of this active inhibition that in fact enables the organism to act under, to 'exploit' the surface-reward association policy, more so than an incremental strengthening of this association itself. Indeed, we see increased exploratory behaviors precisely at moments when we might intuitively expect active inhibition to be lower, such as upon initial introduction into a familiar environment (i.e. beginning of session) or, significantly more so, upon initial introduction into a novel one. Moreover, we identify intra-session time points of significant exploratory behavior precisely with those trials where the population probability of choosing the *un*rewarded surface reaches levels that cannot be accounted for either by previous policy exploitation performances or by purely random choice distribution patterns.

A cognitive interpretation of this is that, just as exploration in the T- and Y-maze is shaped to manifest as spatial alternation by the *physical* constraints of the apparatus, so in our tactile discrimination protocol in the radial maze, what it means to explore is shaped, behaviorally speaking, by prior *cognitive* constraints arising from acquisition of the surface-based state-action policy: to "explore" in the tactile discrimination task is to pointedly visit the *un*rewarded surface. Exploring in this interpretation is not just something which *might* occur in states where the animal makes a decision at random instead of exploiting the optimal reward policy it has nevertheless formed (though this behavior may also sometimes happen). Rather, once the optimal reward policy has been internalized, this appears to constitute a cognitive constraint that shapes exploration to manifest *actively* as a transgression of the policy in moments when we might expect active inhibition to be lowest/not yet engaged. Furthermore, if novelty does indeed boost exploratory behavior (Farahbakhsh & Siciliano, 2021; Lustberg et al., 2020; Park et al., 2021), and if exploration is, as we have just suggested, actively directed towards transgressing the internalized policy,

then this could explain why classical rule reversal protocols have been shown to be more effective when the reversal occurs in a novel environment rather than in the same one where the initial reward-association rule was acquired (McDonald et al., 2004). Similarly, the sheer strength of the exploratory drive elicited by the radial maze apparatus (putatively related to its much larger surface area as compared to a classical T- or Y-maze) gives rise to extremely slow increases in stimulus-response exploitatory behavior, despite the rewarded surface being, to borrow terms from the famous Rescorla-Wagner model of learning, both a reliable and salient predictor (Rescorla & Wagner, 1972).

Precisely what our "indoctrination-like" protocol reveals is that how mice actually revise their innate exploratory state-action policy confounds a view where repeated positive reinforcement simply increases the vigor of the target response. Indeed, such robust active behavioral tendencies make it difficult to see how exploratory behavior could be satisfactorily accounted for simply by increasing the probability of choosing an action at random when in certain states. While the animal behavior literature does also provide a theorization in which reinforcement is taken to be at least as much a case of non-reinforced spontaneous behaviors becoming extinguished over time (Staddon & Simmelhag, 1971), if environmentally elicited *active* exploration requires not extinction but rather ongoing and active *inhibition*, then this requires a different conceptualization again. Furthermore, since the exploratory behavior in our paradigm does appear to be active, as opposed to random, this complicates interpretation of how the mice themselves will interpret a no-reward outcome following an exploratory action. As will be shown and discussed, we have good reason to believe that if there is reward-prediction on exploratory trials, then it is of a measurably different quality to the reward-prediction on exploitatory trials, and this makes it difficult to know to what extent it makes sense to speak of a "reward-prediction *error*" (terminology again borrowed from the Rescorla-Wagner model of learning) when the outcome of an exploratory decision is indeed no-reward.

We might advance that it is fundamental to their evolved nature for opportunistic species, such as mice, rats, humans, and others, to maintain the capacity for vigorous exploratory behavior even after extended periods spent in environments the organism has been able to reliably exploit. And while this is no guarantee that integrating active exploration would therefore be an optimal strategy for reinforcement learning and artificial intelligence, it is interesting to note both that the question of efficient exploration is still deemed to be wide open in several areas of the discipline and that active exploration approaches are one of the avenues currently being pursued in this regard (Khamassi et al., 2017; Ménard et al., 2020; Shyam et al., 2019), alongside approaches which make exploration *intrinsically* (as opposed to just environmentally) reinforcing for the agent (Oudeyer et al., 2007; Schäfer et al., 2022; Singh et al., 2005). As we shall later see in Part 1, if indoctrination is to have meaning then it is precisely in the sense of *active suppression* of innate exploratory drives, of what in lay terms can be called *natural curiosity*. So then, it is worth asking, firstly, whether merely setting the parameters of a state-action policy to "greedy" (i.e. a minimum exploration, maximal immediate reward seeking policy; see Sutton and Barto, 2014) could ever be a suitable proxy for an "indoctrinated" agent. And, secondly, whether we stand to learn something about human behavior by creating learning algorithms which do actually have the capacity to generate meaningfully "indoctrinated" computational agents.

In light of all these considerations regarding persistent active exploration, perhaps the greatest curiosity of the present investigation is that when we subsequently bring "indoctrinated" mice to revise the tactile state-action policy *back* towards an exploratory mode, we observe highly significant, persistent, multi-faceted, and trial-complexity dependent interference. However, in order to arrive at an understanding of why this interference arises in the way it does when nevertheless reverting to spontaneous exploratory behavior, we must first pass under review the multiplicity of cognitive and neural learning and memory systems this process engages.

2. Multiple learning & memory systems

In the brief presentation of state-action policies above, we traced the origin of the concept back to Thorndike's Law of Effect. Now, in considering the development of the idea of multiple learning and memory systems, the natural starting point happens to reside in one of the earliest and most conceptually sophisticated opponents of Thorndike's purely stimulus-response vision of behavior, namely Edward Chace Tolman (Tolman, 1932). Tolman dared to imagine that we might actually be able to use nevertheless strictly behavioral observation to infer things that were happening inside the living "black box" situated between stimuli and responses, i.e. the mind-brain of the behaving organism. From this starting point in Tolman, we will then trace some of the major historical advancements in the idea of multiple learning and memory systems, describing the research landscape in which the behavioral paradigms of the present study were designed and their results interpreted.

2.1 Beyond black box behaviorism.

Tolman's first great innovations in the theory of learning came in his concepts of "latent learning" and "cognitive maps" (Tolman, 1948). Crucially, neither of these concepts were anything that could be accounted for by the stimulus-response/reinforcement learning theories of Tolman's predecessors and contemporaries, such as Thorndike, Skinner, Watson, Hull, etc. Very simply, all the while maintaining an observationally-grounded behaviorist methodology, what Tolman did was demonstrate that learning could occur even in the absence of reinforcement. Let us take a moment to look at how he approached this demonstration experimentally.

Tolman conceived of a simple yet elegant experiment in which he ran three groups of rats in what he called a 6-unit alley T-maze, essentially comprised of three interconnected T-mazes, with a start-point and an end-point (where a food reward could be optionally

placed) separated from each other by six choice-points. The first group of rats found a food reward at the end point starting from day 1. The second group found a food reward there starting only from day 7, and the third group starting from day 3. In other words, the first group was reinforced for completing the maze from the outset, the other two groups only from a delayed timepoint onwards, meaning their initial runs in the maze were *not* reinforced. Counting the number of wrong turns each rat made before arriving at the end-point, Tolman observed that the first group learned gradually and incrementally, session by session, to make less errors and arrive more directly to the point of reinforcement. Importantly, this kind of gradually improving performance towards a reinforced goal had already been provided with explanations using pure stimulus-response/reinforcer type hypotheses: the food reward is a primary reinforcer, the last maze-turn to be taken before reaching it a secondary reinforcer, the second-last maze-turn another secondary reinforcer contingent on the last one, etc. During initial runs, groups 2 and 3 did not show any such gradual "improvement" in their maze navigation in the non-reinforced sessions, since they were not motivated to reach any particular point more than any other. However, following their first reinforcement, in sessions 7 and 3, respectively, they did not subsequently demonstrate gradual and incremental performance improvement in the way group 1 had. Instead, their performance improved by a significant leap between the first reinforced session to the next, and this leap was all the more significant in group 2, first reinforced in session 7, than in group 3, first reinforced in session 3. These leaps in performance confounded simple stimulus-response/reinforcer type explanations. What Tolman instead concluded is that the rats, simply by navigating the maze without any reward objective, were nevertheless learning something about it. This he called *latent* learning, in the sense that there was learning occurring on the cognitive level which had not yet been provided with an occasion to be observably manifest. This occasion was then provided by the introduction at the end point of the maze, during a later session, of a positive reinforcer. In other words, the non-reinforced rats were forming some kind of cognitive map as they

navigated the maze, and this fact became observable as soon as the rats were provided with the environmental motivation to recall that map in order to arrive as directly as possible to a specific point in the territory. From this explanation, it is clear in what sense "latent learning" and the notion of "cognitive maps" go hand in hand in Tolman's learning theory.

In this way, Tolman laid the groundwork not only for consideration of multiple distinct forms of learning but also for how these may interact during memory-based recall. Indeed, it was precisely by designing an experiment with the ability to show how classical stimulus-response reinforcement learning and latent cognitive map learning *interact* that Tolman was able to *disentangle* the presence of both. As we shall see below, this inspired later researchers to adopt similar approaches and to similar powerful effect. Tolman himself would also continue to complexify our understanding of how learning occurs, explicitly pursuing a pluralistic vision throughout his career, with articles such as "There is more than one kind of learning" (Tolman, 1949). Through this work, he was instrumental in the emergence of the cognitive sciences, before anything was known about what neural functions might be responsible for the various kinds of learning he had nevertheless observed through subtle variation of experimentally elicited behaviors.

2.2 Multiple brain systems for learning and memory.

2.2.1 Cortico-hippocampal episodic memory.

Later research, some of it Nobel prize-winning, employed *in vivo* electrophysiological recordings in rats to neurophysiologically situate the cognitive maps Tolman had inferred only from behavior within the hippocampus (O'Keefe, 1976; O'Keefe & Dostrovsky, 1971; O'Keefe & Nadel, 1978). O'Keefe and Nadel further advanced that the hippocampus contributed to memory by mapping experiences not only spatially, i.e. according to *where* they had happened, but also temporally, i.e. according to *when* they had happened. This spatiotemporal interpretation of hippocampal memory function represented a fertile

proximity with then still recent work in human psychology from Elvin Tulving, who as a complement to "semantic memory" (i.e. memory of abstract facts, "The earth is 4 billion years old," "Gandalf is a wizard," etc.) had theorized the concept of "episodic memory," defined by him as "information about temporally dated episodes or events, and temporal-spatial relations among these events" (Tulving, 1972). Archetypal examples of episodic memory can therefore be thought of as (honest) answers to any question of type "*Where* were you *when* X happened?"

Along with the earlier famous case of patient H.M., in whom severe and lasting episodic amnesia was produced by therapeutic resection of the hippocampus-containing medial temporal lobe (Scoville & Milner, 1957), these experimental and theoretical advances led to an explosion of research into hippocampal function which continues to the present. Since then, beyond its role in the formation of spatiotemporal episodic memories (Eichenbaum, 2017a; Ranganath, 2019; Sellami et al., 2017), a vast literature has demonstrated that the hippocampus is also centrally involved in, for example, the formation of associations and subsequent relational memory (Busquets-Garcia et al., 2018; Cohen & Eichenbaum, 1993; Eichenbaum, 2010; Konkel & Cohen, 2009), as well as recollection *per se* (Hirsh, 1974; Hirsh et al., 1978; Ranganath et al., 2004).

The enormous experimental and theoretical contribution Howard Eichenbaum in particular made to our understanding of memory function throughout his long career insisted on the need to complexify our vision, not only of hippocampal function beyond the strictly spatiotemporal, but also of memory itself beyond only the hippocampal formation (Byrne, 2008; Eichenbaum, 2010, 2016, 2017b; Eichenbaum & Cohen, 2001). Interestingly, continuing in the footsteps of Tolman, one of Eichenbaum's major motivations was to show that a limit asserted by one of his predecessors was not justified. In this case, the limit in question was described by Tulving himself, in his claim that episodic memory was an exclusively human cognitive function. Eichenbaum, driven by the

conviction that animal models were the most fertile territory available for gaining deep understanding of general brain function, set out to challenge this claim by experimentally demonstrating episodic memory function in rats (Ranganath, 2019). In a nutshell, the global theoretical approach consists in tying cognitive memory function to neurophysiological brain function to such an extent that where we observe the latter to be sufficiently comparable across species then we should expect to observe the former, provided the presence of appropriate environmental conditions for the animal to interact with. Indeed, Eichenbaum and Cohen (2001) draw a twofold conclusion with respect to the relationship between general brain function and memory: first, memory is "a consequence of the fundamental plasticity of the brain" and is thereby "tied to ongoing information processing in the brain"; secondly, since information processing is organized across "several functional systems," thus "there are multiple forms of memory that have distinct psychological and information processing characteristics, composing multiple, functionally and anatomically distinct memory systems" (Byrne, 2008). In short, the hypothesis here is that if memory is indeed based on an essentially ubiquitous neuronal phenomenon such as brain plasticity then it can *only* be multiple both in neural basis and cognitive function.

2.2.2 Cortico-striatal procedural memory.

Relative to the above discussions of early behaviorist interpretations of learning, the idea that different observable forms of learning and memory would be associated with distinct neural functions also led to experimental demonstrations that procedural or habitual memory (corresponding most closely to the kind of incremental stimulus-response learning Thorndike, Skinner, etc., imagined could explain all animal behavior) relied on cortico-striatal rather than hippocampal function (Balleine & O'Doherty, 2010; Cohen et al., 1997; Eichenbaum, 2010; Gremel & Costa, 2013; McDonald & White, 1993; M. Packard

et al., 1989; M. G. Packard & McGaugh, 1996). Anecdotally, the case of patient H.M. is also instructive in this regard, since he was perfectly capable, through practice, of learning and improving a new motor skill, even though from lesson to lesson he would have no recollection of the previous episode of instruction (Corkin, 1968; Eichenbaum, 2013). This striatum-mediated procedural learning and memory is the primary focus of Part 1 of the present work, in which we develop the "indoctrination-like" anti-exploratory protocol described above.

2.2.3 Amygdalar affective/emotional memory.

It also led to the further dissociation of an *affective* memory system, distinct from both cortico-hippocampal declarative memory and cortico-striatal procedural memory, this time strongly associated with amygdalar function (Aggleton & Mishkin, 1986; Eichenbaum, 2010; LeDoux, 1993; McDonald et al., 2004; McDonald & Hong, 2004; McDonald & White, 1993; White & McDonald, 2002). It is through the affective memory system that an emotional dimension is brought to learning and recollection. Huge research efforts over the last 30 years or so have demonstrated how this emotional dimension contributes (most often, though not always, beneficially) to behavior and cognition. Examples are the capacity for rapid behavioral threat response that bypasses slower cortical processing (LeDoux, 1990, 1992), somatic sensitivity to choice-contingent reward losses too complex for explicit cortical calculation (Bechara & Damasio, 2005), or the fundamental appetitive and motivational "seeking" drive to explore the world at all (Panksepp, 1998).

Crucial to all of these discoveries was innovative behavioral experimental design. In the cited works from Packard, McDonald, White, and Hong, for example, experimental design capable of demonstrating multiple memory system dissociation relied heavily on the numerous modular possibilities offered by the 8-arm radial maze apparatus, the same piece

of experimental equipment chosen by us in the present study in order to investigate context-based interactions between up to four distinct memory systems.

2.2.4 Working memory and cognitive control.

Which brings us to the last memory system to be discussed here, last but perhaps most well-known, by name at least; working memory. To begin, we can return to the example of patient H.M., in whom loss of the medial temporal lobe had given rise to a total incapacity to store novel facts or events in long-term memory. Despite this extreme functional loss, (Scoville & Milner, 1957) were able to observe that patient H.M. had nevertheless retained the ability to, for example, repeat back a string of digits he had just had spoken to him, indicating that whichever brain function underpinned this particular memory capacity was not fundamentally reliant upon the hippocampus. The memory system patient H.M. could rely on to do this is now commonly referred to as "working memory," after seminal work notably by Alan Baddeley beginning in the 1970s (Baddeley, 1992; Baddeley & Hitch, 1974). Baddeley insisted on the fact that this mnemonic function was not merely a passive short-term store but was rather active, context-dependent, and manipulable (hence *working*). From his earliest (human) experimental and theoretical texts on the subject, he linked working memory function directly to retrieval. In fact, his final major publication prior to shifting to the label "working memory" is entitled "Retrieval rules and semantic coding in short-term memory" (Baddeley, 1972). Moreover, in the same text, retrieval itself is linked to the possibility of *intrusions*, i.e. retrieved cognitive content which is either not relevant to the task at hand, such as retrieving a letter in a digit-based task, or which is relevant but mistaken, such as retrieving the wrong digit. It was also Baddeley who began employing the now familiar term "executive" to describe certain functions of working memory, including retrieval and allocation of attention.

Given that the label "working memory" applies to such a wide range of cognitive functions which may be engaged in various combinations as an organism interacts with its environment, it follows, from the words of Eichenbaum and Cohen quoted above, that these functions certainly also correspond to distinct neural circuits. The first work in this direction actually predates Baddeley's theorization of working memory as such and was carried out in monkeys by C. Jacobsen. He observed that monkeys with prefrontal cortex (PFC) ablation displayed a deficit in a delayed-response task (Jacobsen, 1936) of the type that would later be recognized as a working memory task. As mentioned, from the earliest theoretical discussions of working memory in humans, it has been associated with cognitive control, retrieval, and intrusions. In this latter respect, the last decade has seen a significant increase in research into memory retrieval-related *active* or *adaptive forgetting* of interfering or intrusive cognitive content, which underpins precise memory recall (Anderson & Hulbert, 2021; Bekinschtein et al., 2018, 2018; Wimber et al., 2015). This research explicitly ties this active forgetting function to working memory and its central neural mechanism has been identified with top-down PFC-mediated inhibitory control of hippocampal activity (Anderson & Floresco, 2021).

In laboratory rodents, working memory tasks come in several varieties (Dudchenko, 2004), including the classical radial maze working memory task (Olton & Samuelson, 1976) and the T- or Y-maze working memory task (Deacon & Rawlins, 2006; Shoji et al., 2012; Wenk, 2001). The everyday-like memory (Al Abed et al., 2016) and everyday-like rule revision radial maze tasks we employ in Part 2 of this study imply both working memory and active forgetting dimensions. Indeed, a disadvantage of the T- or Y-maze spatial working memory tasks may reside precisely in the fact that they do not provide the occasion for active forgetting to be engaged, since there is no, or very little, context-relevant cognitive content which could cause significant interference. On this point, in Part 2 of the present study, we draw attention to the fact that one of the transgenic mouse lines we test in the everyday-

like memory task displays an extreme deficit in its working memory dimension, and yet the same mouse line has previously been described as having no deficit in working memory on the basis of the simpler T-maze protocol (Albayram et al., 2016). Based on this discrepancy, we advance that the mouse line in question is impaired specifically in its capacity for active forgetting. Yet since, in real world terms, active forgetting is precisely part of our "everyday-like" working memory demands, this raises the question of the extent to which an animal task which does not have a prominent active forgetting component should be described as a model of something as multifaceted as working memory.

2.3 Different systems, different revisions.

We have now briefly reviewed four different learning and memory systems and their respective putative neural bases: 1) cortico-hippocampal spatiotemporal episodic learning and memory; 2) cortico-striatal procedural and habitual learning and memory; 3) amygdalar affective or emotional learning and memory, and; 4) prefrontal cortex-mediated working memory, incorporating cognitive control and active forgetting. In Part 2 of this study, we will see how the everyday-like rule revision paradigm differentially engages all four of these systems during both the pre- and post-choice phases of decision-making and also as a function of trial complexity. This will enable us to qualify, if not yet precisely quantify, their respective contributions to cognition under conditions of novel environment state-action policy revision. Notably, it will become clear that there is a significant and observable difference in the rates of policy/rule revision between each memory system, with working memory updating the fastest and procedural memory the slowest, a certain subtly persistent affective memory phenomenon notwithstanding. The translational relevance of these differences is wholly contained in the term *everyday-like*, since we maintain that real world learning, memory, and state-action policy revision

typically occur in humans under conditions where all four of these cognitive and affective dimensions are present.

More and more, however, our everyday lives also imply an obligation (a social one at least) to reason about increasingly complex subjects, such as epidemiology, virology, immunology, climate science, international diplomacy and economics, etc. Reflecting the work of Damasio mentioned above, in such complex epistemic conditions the cortico-hippocampal capacity to weigh up and comparatively evaluate all available relevant factors is rapidly exhausted, figuratively and perhaps literally overcome with noise, with the result that the agent instead responds using affective and/or procedural learning and memory. In this regard, observing and interpreting which of these four memory systems are or are not significantly impacted by trial complexity in our everyday-like rule revision paradigm is one of the most powerful experimental innovations presented here. For example, we will see that cortico-hippocampal memory performance is significantly impacted by trial complexity whereas the post-choice signals of affective memory are not. We believe our observation of just such discrepancies provides the most persuasive evidence that the behaviors elicited by our paradigm are eminently comparable with that phenomenon which has long been described in humans and is now commonly referred to as 'myside' confirmation bias.

3. Confirmation bias

Nickerson (1998) explains that the most psychologically interesting dimension of biased evidence seeking and evaluation is the unconscious kind. Indeed, related to the brief discussion of active forgetting above, in a certain sense we might even describe the general mechanism of confirmation bias as the non-recognition (because of high uncertainty), and consequent non-inhibition, of intrusive or interfering cognitive content.

3.1 Nomenclature: 'myside' or choice?

In the more recent literature on confirmation bias, a new nomenclature has emerged which subdivides the concept into two quite distinct, though putatively interacting, cognitive phenomena: 1) 'Myside' bias, or how an agent over-values novel information which confirms previously internalized beliefs or other state-action policies (Mercier & Sperber, 2017; Stanovich et al., 2013; Stanovich & West, 2007), and; 2) choice-confirmation bias, whose effects are more immediately the product of favoring repetition of choices which have just led to better than expected outcomes (Chierchia et al., 2021; Palminteri, 2021; Palminteri et al., 2017). Myside bias corresponds to the object of study found in the classical literature review "Confirmation Bias: A Ubiquitous Phenomenon In Many Guises" (Nickerson, 1998), and is the object of investigation of the present study. In view of this title, it is fitting that choice-confirmation bias has emerged as a means of isolating one such guise in order to study it with greater precision. So, although it is not the central object of our own investigation, it was important to us to embrace the research potential such conceptual and functional clarification provides, which is why we do open the door to a choice-confirmation bias analysis of our findings in Part 2, highlighting its potential for dedicated future research. From this point on in the present text, however, "myside bias" and "confirmation bias" will be used interchangeably, with "choice-confirmation bias" specified as such where mentioned.

3.2 Biased notions about myside bias.

3.2.1 "The smarter you are, the less biased you'll be."

One of the most natural things to imagine, something which plays out hundreds of thousands of times daily on social media and elsewhere, is that someone in whom we can very easily observe the myside bias must therefore be severely lacking in intelligence, since otherwise they would surely see it themselves: We are right, they are wrong; they don't change their mind when we present them with arguments we have found to be convincing, therefore they must be dumb. However, recent work has begun to empirically demonstrate that strength of myside bias is actually independent of cognitive ability and does not correlate to standard measures of general intelligence (Macpherson & Stanovich, 2007; Stanovich et al., 2013; Stanovich & West, 2007). Although this seems counter-intuitive, it should not be surprising, since clear bases for drawing this same conclusion are present throughout Nickerson's classical review on confirmation bias. For example, Nickerson tells us that even Francis Bacon, describing the psychological mechanism we now refer to as confirmation bias, stated that philosophers and scientists did not escape the tendency (Nickerson, 1998). We might also refer to the infamous so-called "Nobel disease" or "Nobelitis", being a trend that has been noticed for Nobel prize-winners (hence, *de facto*, presumably very intelligent individuals) to seemingly disproportionately go on to be convinced by pseudo-science or worse, despite mountains of evidence indicating their lack of justification for doing so (Diamandis, 2013). The example of Nobel disease can serve us as more than an interesting curiosity, however. Importantly, the majority of occurrences of it happen when the scientist in question suddenly takes an interest in a domain *outside* of the one s/he won a Nobel prize for. In this sense, their confirmation bias with respect to evidence that is disconfirmatory towards their new pet position is occurring *beyond* the epistemic zone in which they enjoy the highest level of certainty, i.e. their domain of expertise. If we parallel this to the case of the average person, an individual could score very

highly in general intelligence tests, have a very high IQ, yet not at all be educationally equipped to understand the complex ins and outs of, say, climate science or molecular biology. For such an individual, these would represent domains of high uncertainty, regardless of their level of cognitive ability, or general intelligence, or indeed impression of their own level of understanding (Sloman & Fernbach, 2017). Whether the uncertainty be due to technical or to moral complexity, it is in such individual-specific high uncertainty domains we should expect to observe most confirmation bias, even more so if the domain also implies a strong affective dimension for the individual, such as politics (Kaplan et al., 2016; Stanovich, 2021). Indeed, politics is a domain where people tend to stake out broad-stroke positions such as 'left' or 'conservative' or 'libertarian', in a way that is highly susceptible to give rise to selective evidence seeking and biased weighting of information relevant to politically charged, morally or technically complex issues (e.g. trans rights or climate science, respectively). This is a phenomenon that is only aggravated further in the algorithmic world of social media (Cinelli et al., 2021; Lazer et al., 2018). Force of habit and affect should not be underestimated in these situations of high uncertainty that go beyond conscious cognitive ability. Indeed, accurate estimation of their contribution may help explain why strength of myside bias, if it does indeed primarily arise from procedural and affective memory, is not correlated to measures of general intelligence, typically focused on cortico-hippocampal cognitive functions. Indeed, as Stanovich remarks, despite this clear dissociation, no standard measures of general intelligence yet assess the cognitive ability to, for example, overcome confirmation bias (Stanovich et al., 2013).

3.2.2 "Of course animals don't [or do] display myside bias!"

Soon after design and pilot validation of the everyday-like rule revision task as an animal model for myside confirmation bias, I happened to be reading an article in *The New Yorker* (Kolbert, 2017) discussing Hugo Mercier and Dan Sperber's then new book *The Enigma of*

Reason (Mercier & Sperber, 2017). The article laid out the fundamentals of Mercier and Sperber's theory of the evolution of reason in humankind: "Reason developed not to enable us to solve abstract, logical problems or even to help us draw conclusions from unfamiliar data; rather, it developed to resolve the problems posed by living in collaborative groups." Part of this evolutionary scale solution to social problems, the authors advance, is specifically *persuasive* reason, the capacity to weave together arguments with the capacity to convince others to do what we think is best for the group. In such socio-epistemic conditions, developing a stronger cognitive capacity for persuasive reasoning than for strictly factual or critical reasoning would carry a reproductive advantage, particularly in asserting oneself into a position of authority over the group: "There was little advantage in reasoning clearly, while much was to be gained from winning arguments" (Kolbert, 2017). However, Mercier and Sperber go a step further in their claim that confirmation bias must have first evolved in humans, in whom they say it confers a selective advantage; they also explicitly claim that non-human animals could *not* have evolved the cognitive capacity for confirmation bias, because in animals it would threaten survival. As quoted in the *New Yorker* article: "Imagine, Mercier and Sperber suggest, a mouse that thinks the way we do. Such a mouse, 'bent on confirming its belief that there are no cats around,' would soon be dinner." First impressions upon reading this should be that the illustration used is not analogous to what we, to what *they* label 'myside' bias in humans. If a human were reasoning analogously to this hypothetical mouse, we would most likely label it a psychosis, not confirmation bias. Yet, in their book itself, the authors go further still, claiming "Unsurprisingly, then, no confirmation bias emerges from studies of animal behavior." On the one hand, this is, or at least was, trivially true. On the other hand, were it a statement made about human rather than animal research, the authors would surely have concluded that this was a hypothesis which demanded direct empirical testing instead of *a priori* dismissal. As such, this dismissal itself could be interpreted as confirmation bias at work, in precisely the sense described by Francis Bacon:

"The human understanding when it has once adopted an opinion [...] draws all things else to support and agree with it" (Bacon, 1620; Nickerson, 1998).

A fundamental, albeit neglected implication of confirmation bias is that it should be just as likely to underpin "correct" as "incorrect" responses. In the school of philosophy known as "virtue epistemology" there is much discussion of something called "epistemic luck," which is when an agent believes something that is correct but by virtue of luck rather than of "proper" thinking (Pritchard, 2005; Turri & Sosa, 2013). The implication of this is that, all other things being equal, both a person who disagrees with us and a person who agrees with us on a given question may be equally likely to have arrived at their respective positions via the effects of myside bias. So, it should not have been surprising when I later still stumbled across a presumption in animal behavior specialist Jaak Panksepp's work that confirmation bias was something which we should of course expect to see manifest in the behavior of rats, for example (Panksepp, 1998). In short, in the absence of actual empirical testing of the question through specifically designed experiments (Popper, 1935), and although they reach opposing conclusions, both Panksepp and Mercier and Sperber were likely reasoning to a comparable extent under the action of myside bias.

Various other facets of myside confirmation bias will be discussed again at length in Part 2 of this work. Naturally, I have attempted to temper the influence of my own myside bias at every step of this investigation; conception, experimentation, data collection and analysis, and interpretation. However, the greatest safeguard against confirmation bias that we possess as a species, and on this point I agree with Mercier and Sperber, resides in the good faith confrontation of our own beliefs and convictions with those of others in a spirit of reciprocal learning and progress. On which note, I invite the reader to study the content of this PhD project with a mind as critical as it is open, and look forward to the good faith confrontations to follow.

References:

Aggleton, J. P., & Mishkin, M. (1986). The Amygdala: Sensory Gateway to the Emotions. In R. Plutchik & H. Kellerman (Eds.), *Biological Foundations of Emotion* (pp. 281–299). Academic Press. https://doi.org/10.1016/B978-0-12-558703-7.50018-8

Al Abed, A. S., Sellami, A., Brayda-Bruno, L., Lamothe, V., Noguès, X., Potier, M., Bennetau-Pelissero, C., & Marighetto, A. (2016). Estradiol enhances retention but not organization of hippocampus-dependent memory in intact male mice. *Psychoneuroendocrinology*, *69*, 77–89. https://doi.org/10.1016/j.psyneuen.2016.03.014

Albayram, O., Passlick, S., Bilkei-Gorzo, A., Zimmer, A., & Steinhäuser, C. (2016). Physiological impact of CB1 receptor expression by hippocampal GABAergic interneurons. *Pflügers Archiv - European Journal of Physiology*, *468*(4), 727–737. https://doi.org/10.1007/s00424-015-1782-5

Anderson, M. C., & Floresco, S. B. (2021). Prefrontal-hippocampal interactions supporting the extinction of emotional memories: The retrieval stopping model. *Neuropsychopharmacology*, 1–16. https://doi.org/10.1038/s41386-021-01131-1

Anderson, M. C., & Hulbert, J. C. (2021). Active Forgetting: Adaptation of Memory by Prefrontal Control. *Annual Review of Psychology*, *72*(1), 1–36. https://doi.org/10.1146/annurev-psych-072720-094140

B F Skinner. (1938). *The Behavior Of Organisms An Experimental Analysis*. http://archive.org/details/in.ernet.dli.2015.191112

Bacon, F. (1620). The New Organon: Or True Directions Concerning the Interpretation of Nature. *The New Organon*, 134.

Baddeley, A. D. (1972). Retrieval rules and semantic coding in short-term memory. *Psychological Bulletin*, *78*(5), 379–385. https://doi.org/10.1037/h0033477

Baddeley, A. D. (1992). Working Memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Balleine, B. W., & O'Doherty, J. P. (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, *52*(2), 336–372. https://doi.org/10.1016/j.geb.2004.06.010

Bekinschtein, P., Weisstaub, N. V., Gallo, F., Renner, M., & Anderson, M. C. (2018). A retrieval-specific mechanism of adaptive forgetting in the mammalian brain. *Nature Communications*, *9*(1), 4660. https://doi.org/10.1038/s41467-018-07128-7

Busquets-Garcia, A., Oliveira da Cruz, J. F., Terral, G., Pagano Zottola, A. C., Soria-Gómez, E., Contini, A., Martin, H., Redon, B., Varilh, M., Ioannidou, C., Drago, F., Massa, F., Fioramonti, X., Trifilieff, P., Ferreira, G., & Marsicano, G. (2018). Hippocampal CB1 Receptors Control Incidental Associations. *Neuron*, *99*(6), 1247-1259.e7. https://doi.org/10.1016/j.neuron.2018.08.014

Byrne, J. H. (Ed.). (2008). *Learning and memory: A comprehensive reference* (1st ed). Elsevier.

Chierchia, G., Soukupová, M., Kilford, E. J., Griffin, C., Leung, J. T., Blakemore, S.-J., & Palminteri, S. (2021). *Choice-confirmation bias in reinforcement learning changes with age during adolescence*. PsyArXiv. https://doi.org/10.31234/osf.io/xvzwb

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System*. MIT Press.

Cohen, N. J., Poldrack, R. A., & Eichenbaum, H. (1997). Memory for Items and Memory for Relations in the Procedural/Declarative Memory Framework. *Memory*, *5*(1–2), 131–178. https://doi.org/10.1080/741941149

Corkin, S. (1968). Acquisition of motor skill after bilateral medial temporal-lobe excision. *Neuropsychologia*, *6*(3), 255–265. https://doi.org/10.1016/0028-3932(68)90024-9

Deacon, R. M. J., & Rawlins, J. N. P. (2006). T-maze alternation in the rodent. *Nature Protocols*, *1*(1), 7–12. https://doi.org/10.1038/nprot.2006.2

Dember, W. N., & Richman, C. L. (1989). *Spontaneous Alternation Behavior*. Springer New York. https://doi.org/10.1007/978-1-4613-8879-1

Diamandis, E. P. (2013). Nobelitis: A common disease among Nobel laureates? *Clinical Chemistry and Laboratory Medicine*, *51*(8), 1573–1574. https://doi.org/10.1515/cclm-2013-0273

Dudchenko, P. A. (2004). An overview of the tasks used to test working memory in rodents. *Neuroscience & Biobehavioral Reviews*, *28*(7), 699–709. https://doi.org/10.1016/j.neubiorev.2004.09.002

Eichenbaum, H. (2010). Memory systems. *WIREs Cognitive Science*, *1*(4), 478–490. https://doi.org/10.1002/wcs.49

Eichenbaum, H. (2013). What H.M. taught us. *Journal of Cognitive Neuroscience*, *25*(1), 14–21. https://doi.org/10.1162/jocn_a_00285

Eichenbaum, H. (2016). What Versus Where: Non-spatial Aspects of Memory Representation by the Hippocampus. In R. E. Clark & S. J. Martin (Eds.), *Behavioral Neuroscience of Learning and Memory* (Vol. 37, pp. 101–117). Springer International Publishing. https://doi.org/10.1007/7854_2016_450

Eichenbaum, H. (2017a). On the Integration of Space, Time, and Memory. *Neuron*, *95*(5), 1007–1018. https://doi.org/10.1016/j.neuron.2017.06.036

Eichenbaum, H. (2017b). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558. https://doi.org/10.1038/nrn.2017.74

Eichenbaum, H., & Cohen, N. J. (2001). *From Conditioning to Conscious Recollection: Memory systems of the brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195178043.001.0001

Farahbakhsh, Z. Z., & Siciliano, C. A. (2021). Neurobiology of novelty seeking. *Science*, *372*(6543), 684–685. https://doi.org/10.1126/science.abi7270

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

Gremel, C. M., & Costa, R. M. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature Communications*, *4*(1), 2264. https://doi.org/10.1038/ncomms3264

Habedank, A., Kahnau, P., & Lewejohann, L. (2021). Alternate without alternative: Neither preference nor learning explains behaviour of C57BL/6J mice in the T-maze. *Behaviour*, *158*(7), 625–662. https://doi.org/10.1163/1568539X-bja10085

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology*, *12*(4), 421–444. https://doi.org/10.1016/S0091-6773(74)92231-7

Hirsh, R., Leber, B., & Gillman, K. (1978). Fornix fibers and motivational states as controllers of behavior: A study stimulated by the contextual retrieval theory. *Behavioral Biology*, *22*(4), 463–478. https://doi.org/10.1016/S0091-6773(78)92583-X

Inglis, I. R., Langton, S., Forkman, B., & Lazarus, J. (2001). An information primacy model of exploratory and foraging behaviour. *Animal Behaviour*, *62*(3), 543–557. https://doi.org/10.1006/anbe.2001.1780

Jacobsen, C. F. (1936). Studies of cerebral function in primates. I. The functions of the frontal association areas in monkeys. *Comparative Psychology Monographs*, *13, 3*, 1–60.

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*(1), 39589. https://doi.org/10.1038/srep39589

Khamassi, M., Velentzas, G., Tsitsimis, T., & Tzafestas, C. (2017). Active Exploration and Parameterized Reinforcement Learning Applied to a Simulated Human-Robot Interaction Task. *2017 First IEEE International Conference on Robotic Computing (IRC)*, 28–35. https://doi.org/10.1109/IRC.2017.33

Kolbert, E. (2017, February 19). Why Facts Don't Change Our Minds. *The New Yorker*. http://www.newyorker.com/magazine/2017/02/27/why-facts-dont-change-our-minds

Konkel, A., & Cohen, N. (2009). Relational memory and the hippocampus: Representations and methods. *Frontiers in Neuroscience*, *3*. https://www.frontiersin.org/article/10.3389/neuro.01.023.2009

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*. https://doi.org/10.1126/science.aao2998

LeDoux, J. E. (1990). Information flow from sensation to emotion: Plasticity in the neural computation of stimulus value. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience* (pp. 3–51). Bradford Books (MIT Press).

LeDoux, J. E. (1992). Brain mechanisms of emotion and emotional learning. *Current Opinion in Neurobiology*, *2*(2), 191–197. https://doi.org/10.1016/0959-4388(92)90011-9

LeDoux, J. E. (1993). Emotional memory systems in the brain. *Behavioural Brain Research*, *58*(1), 69–79. https://doi.org/10.1016/0166-4328(93)90091-4

Lustberg, D., Tillage, R. P., Bai, Y., Pruitt, M., Liles, L. C., & Weinshenker, D. (2020). Noradrenergic circuits in the forebrain control affective responses to novelty. *Psychopharmacology*, *237*(11), 3337–3355. https://doi.org/10.1007/s00213-020-05615-8

Macpherson, R., & Stanovich, K. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, *17*(2), 115–127. https://doi.org/10.1016/j.lindif.2007.05.003

McDonald, R. J., Foong, N., & Hong, N. S. (2004). Incidental information acquired by the amygdala during acquisition of a stimulus-response habit task. *Experimental Brain Research*, *159*(1), 72–83. https://doi.org/10.1007/s00221-004-1934-x

McDonald, R. J., & Hong, N. S. (2004). A dissociation of dorso-lateral striatum and amygdala function on the same stimulus–response habit task. *Neuroscience*, *124*(3), 507–513. https://doi.org/10.1016/j.neuroscience.2003.11.041

McDonald, R. J., & White, N. M. (1993). *A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. - PsycNET*. https://content.apa.org/doiLanding?doi=10.1037%2F0735-7044.107.1.3

Ménard, P., Domingues, O., Jonsson, A., Kaufmann, E., Leurent, E., & Valko, M. (2020). *Fast active learning for pure exploration in reinforcement learning*. 37.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, *135*(5), 601. https://doi.org/10.1037/bne0000471

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, *51*(1), 78–109. https://doi.org/10.1016/0014-4886(76)90055-8

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press ; Oxford University Press.

Olton, D. S., & Samuelson, R. J. (1976). Remembrance of places passed: Spatial memory in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*(2), 97–116. https://doi.org/10.1037/0097-7403.2.2.97

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, *11*(2), 265–286. https://doi.org/10.1109/TEVC.2006.890271

Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, *65*(1), 65–72. https://doi.org/10.1006/nlme.1996.0007

Packard, M., Hirsh, R., & White, N. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *The Journal of Neuroscience*, *9*(5), 1465–1472. https://doi.org/10.1523/JNEUROSCI.09-05-01465.1989

Palminteri, S. (2021). *Choice-confirmation bias and gradual perseveration in human reinforcement learning*. PsyArXiv. https://doi.org/10.31234/osf.io/dpqj6

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.

Park, A. J., Harris, A. Z., Martyniuk, K. M., Chang, C.-Y., Abbas, A. I., Lowes, D. C., Kellendonk, C., Gogos, J. A., & Gordon, J. A. (2021). Reset of hippocampal–prefrontal circuitry facilitates learning. *Nature*, *591*(7851), 615–619. https://doi.org/10.1038/s41586-021-03272-1

Popper, K. R. (1935). *The Logic of Scientific Discovery*. Routledge. http://nukweb.nuk.uni-lj.si/login?url=http://search.ebscohost.com/login.aspx?authtype=ip&direct=true&db=nlebk&AN=143035&site=eds-live&scope=site&lang=sl

Pritchard, D. (2005). *Epistemic Luck*. Oxford University Press. https://doi.org/10.1093/019928038X.001.0001

Ranganath, C. (2019). Time, memory, and the legacy of Howard Eichenbaum. *Hippocampus*, *29*(3), 146–161. https://doi.org/10.1002/hipo.23007

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, *42*(1), 2–13. https://doi.org/10.1016/j.neuropsychologia.2003.07.006

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory: Vol. Vol. 2*.

Richman, C. L., Dember, W. N., & Kim, P. (1986). Spontaneous alternation behavior in animals: A review. *Current Psychological Research & Reviews*, *5*(4), 358–391. https://doi.org/10.1007/BF02686603

Schäfer, L., Christianos, F., Hanna, J. P., & Albrecht, S. V. (2022). Decoupled Reinforcement Learning to Stabilise Intrinsically-Motivated Exploration. *ArXiv:2107.08966 [Cs]*. http://arxiv.org/abs/2107.08966

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

Sellami, A., Al Abed, A. S., Brayda-Bruno, L., Etchamendy, N., Valério, S., Oulé, M., Pantaléon, L., Lamothe, V., Potier, M., Bernard, K., Jabourian, M., Herry, C., Mons, N., Piazza, P.-V., Eichenbaum, H., & Marighetto, A. (2017). Temporal binding function of dorsal CA1 is critical for declarative memory formation. *Proceedings of the National Academy of Sciences*, *114*(38), 10262–10267. https://doi.org/10.1073/pnas.1619657114

Shoji, H., Hagihara, H., Takao, K., Hattori, S., & Miyakawa, T. (2012). T-maze Forced Alternation and Left-right Discrimination Tasks for Assessing Working and Reference Memory in Mice. *JoVE (Journal of Visualized Experiments)*, *60*, e3300. https://doi.org/10.3791/3300

Shyam, P., Jaśkowski, W., & Gomez, F. (2019). Model-Based Active Exploration. *Proceedings of the 36th International Conference on Machine Learning*, *5779–5788*. https://proceedings.mlr.press/v97/shyam19a.html

Singh, S., Barto, A. G., & Chentanez, N. (2005). *Intrinsically Motivated Reinforcement Learning:* Defense Technical Information Center. https://doi.org/10.21236/ADA440280

Sloman, S., & Fernbach, P. (2017). *The Knowledge Illusion: Why We Never Think Alone*. Penguin.

Staddon, J. E., & Simmelhag, V. L. (1971). The "supersitition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psychological Review*, *78*(1), 3–43. https://doi.org/10.1037/h0030305

Stanovich, K. (2021). *The Bias That Divides Us: The Science and Politics of Myside Thinking*. MIT Press.

Stanovich, K., & West, R. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning - THINK REASONING*, *13*, 225–247. https://doi.org/10.1080/13546780600780796

Stanovich, K., West, R., & Toplak, M. (2013). Myside Bias, Rational Thinking, and Intelligence. *Current Directions in Psychological Science*, *22*, 259–264. https://doi.org/10.1177/0963721413480174

Summerfield, C., & Parpart, P. (2022). Normative Principles for Decision-Making in Natural Environments. *Annual Review of Psychology*, *73*(1), 53–77. https://doi.org/10.1146/annurev-psych-020821-104057

Sutton, R. S., & Barto, A. G. (2014). *Reinforcement Learning: An Introduction* (2nd ed.). A Bradford Book.

Thorndike, E. L. (Edward L. (1911). *Animal intelligence; experimental studies*. New York, The Macmillan company. http://archive.org/details/animalintelligen00thor

Tolman, E. C. (1932). *Purposive behavior in animals and men*. Univ of California Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tolman, E. C. (1949). There is more than one kind of learning. *Psychological Review*, *56*(3), 144–155. https://doi.org/10.1037/h0055304

Tulving, E. (1972). Episodic and semantic memory. In *Organization of memory* (pp. xiii, 423–xiii, 423). Academic Press.

Turri, J., & Sosa, E. (2013). Virtue Epistemology. In B. Kaldis (Ed.), *Encyclopedia of Philosophy and the Social Sciences* (pp. 427–440). Sage Publications.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292. https://doi.org/10.1007/BF00992698

Wenk, G. L. (2001). Assessment of spatial memory using the T maze. *Current Protocols in Neuroscience*, *Chapter 8*, Unit 8.5B. https://doi.org/10.1002/0471142301.ns0805bs04

White, N. M., & McDonald, R. J. (2002). Multiple parallel memory systems in the brain of the rat. *Neurobiology of Learning and Memory*, *77*(2), 125–184. https://doi.org/10.1006/nlme.2001.4008

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, *18*(4), 582–589. https://doi.org/10.1038/nn.3973

* * *

PART I

* * *

"This work was strictly voluntary, but any animal who absented himself from it would have his rations reduced by half."

George Orwell, *Animal Farm* (1945).

Indoctrination as active inhibition of spontaneous exploration: Introduction of a novel mouse model.

Christopher Stevens, Cathy Lacroix, Yamuna Mariani, Giovanni Marsicano, Aline Marighetto.

Abstract

Indoctrination has been defined as any educative process that actively discourages open-mindedness or curiosity. In parallel, few aspects of mouse behavior have been better characterized than their natural curiosity. So, given a recognized lack of knowledge about the cognitive and neurophysiological underpinnings of the phenomenon, we developed a model of indoctrination-like learning in mice. Based around a tactile stimulus-response (S-R) rule in the 8-arm radial maze, our research reveals a decoupling of early rule acquisition from its delayed sustained expression, the latter requiring active inhibition of spontaneous exploration. This reflects the distinction between simple transmission of cognitive content, common to all education, and the extra effort required to transform that content into a curiosity suppressing "doctrine". We began investigating the mechanisms of this decoupling using transgenic and physiological ageing approaches. Deletion of the endocannabinoid type-I receptors (CB_1) from dopamine type-I receptor (D_1) expressing neurons of the forebrain impaired rule expression but not acquisition. This deficit was rescued by viral re-expression of CB_1 specifically in GABAergic neurons of the striatal direct pathway of D_1-CB_1-KO mice. In aged mice also, rule expression but not acquisition was impaired, based upon which we postulate that age-related decrease in cognitive flexibility translates here into reduced capacity to inhibit spontaneous exploratory responses. Combining these results, we suggest that inhibitory cognitive flexibility in the dorsal striatum plays a key role in enabling acquired behavioral strategies to supplant innate ones. We advance, for further investigation in humans, that indoctrination requires, and is perhaps best characterized as a coopting, or "hijacking" of cognitive flexibility as a means to suppressing spontaneous curiosity.

Introduction

Indoctrination has been described as teaching that thwarts "open-mindedness" or curiosity, as education that produces "close-minded" individuals, who in turn are defined as those who are "unable or unwilling to give due regard to reasons that are available for revising their current beliefs" (Callan & Arena, 2009; Taylor, 2017). Such a definition of close-mindedness closely matches that given for the psychological phenomenon of confirmation bias in the go-to literature review on the subject (Nickerson, 1998): an inability or unwillingness to accurately evaluate evidence that casts the truth of a belief or hypothesis into question, while simultaneously over-valuing evidence that confirms said belief or hypothesis. Combined, this implies that indoctrination not only tends to give rise to confirmation bias, but also that where we observe marked confirmation bias we should suspect that the learning processes which have led to it tended towards indoctrination. While indoctrination has most usually been portrayed as undesirable in any democratic educative process (Dewey, 1916), and while public education in developed nations may now be moving in a more "progressive" direction, we nevertheless find influential contemporary historical apologetics for forms of indoctrination in civic and scientific education (Conant, 1948; Crozier et al., 1975; Kuhn, 1977, 1961), as well as the argument that indoctrination of some manner may be an inevitable part of all education (Macmillan, 1983), not to mention an ever-growing mountain of evidence indicating that forms of indoctrination in informal learning contexts, notably via the internet, are as present as ever (Alfano et al., 2018; O'Callaghan et al., 2015). Not surprising then that both terms, indoctrination and confirmation bias, are becoming ever more prominent in public discourse, with the growth of societal concerns such as the widespread dissemination of political and scientific misinformation via traditional and social media networks. Yet despite this growing presence and concern, it has also been remarked that we know little about the cognitive and neural implications or consequences of educative processes, such as indoctrination, which work *against* natural exploratory curiosity (Kaplan et al., 2016; Reynolds & Canna, 2012).

Animal Farm, George Orwell's allegory of political manipulation, may be the first thing to come to mind if asked for an example of animals being indoctrinated. Part of its

power as an allegory derives from a shared assumption that the indoctrination of real world animals is in essence fantastical. However, when it comes to gaining a deeper scientific understanding of the cognition and neurophysiology underpinning any human brain state, firstly, evolution must be accounted for (Cisek & Hayden, 2022) and, secondly, assumptions should be permitted only if accompanied by experimental conditions under which they could, in theory at least, be falsified (Popper, 1935). Thus, when a theory of the evolution of human reasoning leans on an assumption that confirmation bias could only emerge from human brains (Mercier & Sperber, 2017), the coherence of the overarching theory cannot also serve as foundational evidence for the assumption. Nor can we use absence of evidence (e.g. the fact that indoctrination based confirmation bias-like behavior has never been observed in a non-human animal species) as evidence of absence (e.g. the conclusion that indoctrination based confirmation bias-like behavior therefore *cannot* be observed in a non-human animal species, as asserted by Mercier & Sperber, 2017).

Across two papers, of which this is the first, we lay out the results of directly empirically testing and investigating the following hypothesis: non-human animals subjected to real-world analogous indoctrination-like (present paper) and confirmation bias-like (Stevens et al., 2022b) environmental conditions will display behaviors analogous and comparable to human indoctrination and confirmation bias. Beyond the intrinsic interest of observing and investigating this behavior in mice, if validated, the hypothesis would imply that indoctrination and confirmation bias are primarily functions of the particularities of human environments and not primarily functions of the particularities of human brains.

A fundamental characteristic of rodents that makes them so interesting and relevant for the study of behavioral neuroscience consists precisely in their own innate exploratory behavior. In the language of reinforcement learning, this exploratory behavior can be modeled as a state-action policy (Sutton & Barto, 2018) whereby, when in a state S_0, an agent (i.e. mouse) will spontaneously take that action A_n which will move it into a new state S_n representing the greatest available environmental novelty or uncertainty (Frank et al., 2009), e.g. a zone of the environment either not yet or least recently explored. In laboratory conditions, such behavior has been thoroughly investigated in the phenotype

of exploratory spontaneous alternation in the T-maze, Y-maze, and radial maze (Dember & Richman, 1989; Lalonde, 2002; Olton & Samuelson, 1976; Richman et al., 1986). As a result, once we had formalized the core principle of indoctrination as the explicit inhibition of spontaneous exploratory behavior with respect to a cognitive content being taught, we selected an experimental environment in which we knew mouse exploratory behavior was maximal, in order that its suppression would be gradual enough to provide the occasion for both detailed analysis and intervention. Based on previous work from our laboratory (Marighetto et al., 1999), we knew that mice placed in a large 8-arm radial maze (designed to accommodate either mice or rats, hence particularly large relative to mice) would display highly persistent exploratory behavior, even under motivated (food-restricted) conditions where they are being trained to visit and receive a food reward on a fixed subset of arms only. In the language of classical decision-making literature, this reflects what is called the exploration/exploitation trade-off (March, 1991), whereby an agent acts under the opposing influences of the perceived value of exploring new possibilities (e.g. re-visiting arms which have never yet been rewarded) versus the value of repeatedly exploiting known certainties (e.g. re-visiting arms which have always been rewarded). Exploitation implies foregoing exploration and vice-versa, hence "trade-off." In order to provide mice with as unambiguous a "known certainty" to exploit as possible, we conceived of a simple tactile stimulus-response (S-R) rule (Colombo et al., 1990; McDonald & Hong, 2004) of the type referred to as "win-stay" (McDonald & White, 1993; Packard et al., 1989). A "win-stay" rule means that, when rewarded, the agent must subsequently repeat the same action (e.g. press lever, visit arm) in the same place (e.g. left, right, cue location, etc.) in order to be rewarded again. This is in opposition to a "win-shift" rule whereby the agent must switch to the other lever or arm, etc., from the one that was just rewarded in order to be rewarded again.

We divided the arms and central platform of the radial maze according to two surface types, one always rewarded (S1), one never rewarded (S0), and imposed a "win-stay" rule (R1) with respect to S1. To emphasize the tactile dimension, R1 training was conducted in darkness. Across repeated binary choice trials (one trial = one free choice between two contiguous radial maze arms, one S1, one S0), we taught mice that they would receive a food reward only if they visited S1 arms. R1 therefore implied that, in order to maximize food-reward, mice limit their explorations to only half the area of the

environment. Or, in other words, in order to *exploit* R1 successfully, mice had to overcome their innate drive to *explore* continuously, which in a dark but tactile environment they could most reliably achieve by alternating choices between S1 and S0 arms from trial to trial.

S-R learning, such as R1, has been classically associated with striatal structures, specifically in rodents with the dorsolateral striatum (McDonald & White, 1993; Packard et al., 1989). We therefore hypothesized that direct and indirect pathway functions, such as novel action sequence "chunking" (i.e. consolidation of multi-component actions into singular, direct pathway selectable ones, like riding a bicycle or touch-typing) and selection (Graybiel, 1998), would play a role in resistance towards *expression* of R1. This was in part based on a previous study demonstrating that response sequences were more spontaneously "chunked" by rodents under a location "shift" strategy than by rodents under a location "stay" strategy in the T-maze (Cohen et al., 2004). By similar logic, we also reasoned that these same striatal functions would not significantly impact *acquisition* of the S1-reward association, since its acquisition implied no antagonism with other cognitive functions. To investigate this predicted decoupling, we developed behavioral analyses allowing us to evaluate cognitive indicators of R1 acquisition – i.e. decision latency (Carland et al., 2019; Marighetto et al., 2000), post-choice run time (Carland et al., 2019; Dhawale et al., 2021; Dudman & Krakauer, 2016; Marighetto et al., 1999), and deliberative choice revision (Redish, 2016; Tolman, 1939) – as distinct from the primary indicator of R1 expression, i.e. percentage of correct R1 responses.

To more deeply investigate the striatal hypothesis, we employed a genetic approach using the D_1-CB_1-KO transgenic mouse line, in which cannabinoid type-I receptors (CB_1) are conditionally deleted from dopamine type-I receptor (D_1) positive neurons of the forebrain (Monory et al., 2007). Within the striatum, D_1 are expressed on inhibitory medium-spiny GABAergic neurons of the direct pathway. CB_1 expressed on the pre-synaptic element of these neurons exert a retrograde inhibitory modulatory effect, producing a net effect of inhibiting an inhibitory signal. Indeed, deletion of CB_1 receptors has been shown to potentiate the net inhibitory signal of direct pathway neurons (Soria-Gomez et al., 2021). Activation of the direct pathway has been associated with reward coding and "stay" action-selection strategies (Nonomura, 2018; Vicente et al., 2016).

In the present study, however, we reframe these results from the operant behavior literature and hypothesize that the direct pathway "stay" strategy may be better and more broadly interpreted as the repeated ("stay") selection of whichever action is most spontaneously performed in a given task environment, i.e. repeated interaction with one operandum in operant conditions, but repeated alternation of contiguous arms in maze conditions (see Discussion). This interpretation would also imply that the no-reward coding "shift" action-selection associated with D_2-expressing indirect pathway activation (Nonomura, 2018; Vicente et al., 2016) translate in the maze context to "shifting" strategy away from spontaneous (exploratory) alternation to, notably, repeated (exploitatory) selection of only one arm or surface. In this framing, our hypothesis was that persistent direct pathway-mediated selection of the innate exploratory response would be potentiated in D_1-CB_1-KO animals, further disrupting its gradual inhibition via the indirect pathway-mediated no-reward coding "shift" to the novel exploitatory strategy, required for high R1 performance. Crucially, with respect to the distinction made above, if this interpretation were accurate, expression but not acquisition of R1 should be impacted by manipulation of the direct pathway.

Finally, since it is known that both CB_1 and D_1 expression in the central nervous system, including in the striatum, decreases with ageing (Bilkei-Gorzo, 2012; Wang et al., 1998), we hypothesized that resistance to expression of a novel win-stay exploitatory rule would also increase with age. In terms of classical ageing phenotypes, we predicted that, in aged mice, diminishing age-related cognitive flexibility would translate in our radial maze task as more rigid expression, through reduced inhibitory control, of the innate, exploratory strategy, thereby blocking expression of the acquired S-R exploitatory strategy (the latter being the proverbial "new trick" of this scenario). As above, we again predicted that this age-related disruption would be specific to expression and not acquisition of the S1-reward association.

Our results reveal that mice did initially make use of the tactile environmental affordance to guide surface-based exploratory behavior, displaying a robust and significant trend to alternate between S1 and S0 from trial to trial during early sessions. Multiple behavioral analyses revealed that mice began to acquire the S1-reward association as early as the third session of training. However, resistance to sustained expression of S1 exploitatory behavior was vigorous, such that it took up to 15 sessions of training for

some individual mice to reach criterion performance level of 75% correct R1 responses averaged across two sessions. This clear decoupling of rule acquisition from sustained expression demonstrates that, in mice also, simply learning that a certain strategy will be reinforced is not alone sufficient to suppress exploration: a prolonged protocol of indoctrination-like training is required in order to inhibit it. This resistance to R1 expression was more marked in D_1-CB_1-KO mice compared to wildtype littermates, with, however, no observable differences with respect to acquisition. The wildtype R1 expression phenotype was successfully rescued by targeted viral re-expression of CB_1 in the D_1 positive direct pathway of the striatum, through which spontaneous behaviors are putatively selected. Finally, aged mice, classically characterized by a decrease in cognitive flexibility, demonstrated even stronger resistance to expression, but not to acquisition of R1. Though seemingly paradoxical, what we suggest based on these results is that indoctrination requires, or is perhaps best characterized as a coopting (Gould et al., 1979; Gould & Vrba, 1982), a neuronal recycling (Dehaene & Cohen, 2007), or what in addiction studies is called a "hijacking" (Munro, 2015; Schultz, 2016) of cognitive flexibility, using it to inhibit natural curiosity with respect to the cognitive content being learnt.

Materials & Methods

Animals: Young (8 to 12 weeks) C57BL/6J male mice were obtained from Charles River and collectively housed in a standardized animal room (23 °C; lights on 7 AM to 7 PM; four or five mice per cage). Mice from the aged cohort (~18 months) underwent ageing in collective housing on site at the animal facility of the Neurocentre Magendie. D_1-CB_1-KO mice were generated as previously described (Monory et al., 2007; Terzian et al., 2011) by crossing CB_1 floxed mice (Marsicano et al., 2003) with D_1-Cre line mice (Lemberger et al., 2007), in which the Cre recombinase was placed under the control of the D_1 gene (Drd1a). As previously described (Zerucha et al., 2000; Monory et al., 2006), Dlx5/6-Cre mice were crossed with $CB_{1}^{f/f}$ mice to obtain $CB_{1}^{f/f;Dlx5/6\text{-}Cre}$ (here called Dlx-CB_1-KO) and their $CB_{1}^{f/f}$ (WT) littermate controls. 8 to 14-week-old naive male D_1-CB_1-KO and Dlx-CB_1-KO and their respective WT littermates were used. All animals were moved to individual cages 2 weeks before the beginning of experiments.

Food restriction: Five days prior to the first day of training, all animals were placed under a progressive food restriction schedule in order to gradually bring them to 85% to 90% of their baseline free feeding weight (calculated by weighing animals each day for 3 days during ad libitum food access). Individual animal weight and welfare was monitored daily throughout the duration of the experimentation. All experiments were conducted in accordance with European Directive 2010-63-EU and with approval from the Bordeaux University Animal Care and Use Committee CCEA50. All efforts were made to minimize suffering and reduce the number of animals used.

Viruses and surgery: D_1-CB_1-KO mice were anesthetized in an induction box containing 5% Isoflurane (Virbac, France) before being secured in a stereotaxic frame (Model 900, Kopf instruments, CA, USA) in which 1.0% to 1.5% isoflurane was continuously supplied via an anesthetic mask for the duration of the surgery. Animals were injected with local analgesic (Lidocaine/Lidor, 2mg/ml, 100ul per mouse) and opioid analgesic (Buprenorphine/Buprecare, 0.3 mg/ml,100 ul per mouse) at the beginning of surgery. For viral intra-striatal AAV delivery, AAV vectors were injected with the help of a microsyringe (0.25 mL Hamilton syringe with a 30-gauge beveled needle) attached to a pump (UMP3-1, World Precision Instruments, FL, USA). D_1-CB_1-KO mice were injected directly into the striatum (STR) (1 µl per injection site at a rate of 0.5 µl per min, for a total of 8 µl per animal), with the following bilateral coordinates: AP 1.5; ML ± 2; DV –3.5 / –3, and AP –0.5; ML ± 2.6; DV –3.5 / –3. Following virus delivery at each site, the syringe was left in place for 2 minutes (DV –3.5 sites) and 5 minutes (DV –3 sites) before being slowly withdrawn from the brain. 6 mice were injected with pAAV-CAG-flexx-IRES-mCYT (empty control vector) to create the D_1-CB_1-KO-$Str^{-/-}$ group and 6 mice with pAM-CAG-flexx-CB1myc to induce re-expression of the CB_1 receptor gene in the striatum and create the D_1-CB_1-KO-$Str_{+/+}$ group. At the end of surgery, all operated animals were given an anti-inflammatory injection (METACAM, 2mg/ml, 50ul per mouse i.p.). In this experiment, expression was allowed to take place for 4-5 weeks after local infusions. Mice were monitored and weighed daily post-operation for three days and also given one more i.p. injection of Metacam the day after surgery, as described above. All animals rapidly regained their pre-surgery body weight, meaning none needed to be excluded from the experiments.

Radial maze: The apparatus is an 8-armed fully automated radial maze (Imetronic), the surface of which is raised ~100cm off ground level. Access to each arm is from a central platform by means of automated vertically sliding doors. When all doors are closed during behavior, the experimental animal is contained within the central platform, a regular octagon of size 483cm^2 and edge 10cm (i.e. the width of each arm and door). At the distal end of each 50cm length arm is an automated pellet distributor for dispensing food reward. The distributor is set into a slight indent in order to hide its state (i.e. baited or not baited) from the animal. For this study, we produced removable polymer panels which could be placed so as to cover the entire area of the radial maze. The panels are of two distinct tactile types: smooth surfaced panels (similar to the usual surface of the radial maze) and irregular surfaced panels (the finish of which was a uniform but irregular beveled pattern of < 2mm maximum relief). This allowed us to present the radial maze according to various tactile configurations: entirely smooth, entirely irregular, or various combinations of smooth and irregular. Animal movements are detected via video camera and motion detection software (GenCam) using either visible or infra-red light, depending on the experimental conditions (see below). The motion detection software communicates with a second piece of software, POLYRadial (Imetronic), through which pre-programmed sequences of automated radial maze actions are triggered. This program is used for the design and execution of behavioral exercises (sequences of door openings, location of food reward, conditions for opening and closing of doors, etc.). Hence, the exercises are customizable and contingent upon a combination of the detected movements of the animal and automated timed sequences.

Behavior

Habituation: Prior to the first day of tactile discrimination learning, all animals were habituated to the context and functioning of the radial maze apparatus. Food restriction, as described above, began three days before habituation (i.e. five days before training). At the beginning of each habituation session, the animal was placed by the experimenter in the central platform of the radial maze, all 8 doors of which were closed. Once removed to the control room, the experimenter launched the habituation program via the POLYRadial software. The habituation program began by an interval of 10 seconds during which the animal could explore the central platform. Following this, all 8 doors

opened simultaneously, presenting the animal with the opportunity to freely explore the entire surface of the maze. As the animal explored, once it had advanced to the most distal section of a given arm (location of the distributor and food reward) and returned to the central platform, the door of that arm automatically closed behind it, thus preventing further access to that arm during the current session. Thus, once the animal had fully explored all 8-arms, it found itself again contained within the central platform. At this point, a further habituation session could be launched if needed. It was considered that when an animal had recovered and consumed at least 5 out of 8 available food rewards in a single session that it was fully habituated to the relevant functionalities of the apparatus. All animals reached this habituation criterion within an average of 5 sessions, conducted one after the other without removing the animal from the apparatus, thereby also minimizing stress. Since our R1 tactile discrimination task was to be conducted in the dark, thus potentially making it difficult for animals to perceive that doors always opened in contiguous pairs, animals also underwent a second habituation session 24 hours after the first, during which pairs of doors opened simultaneously, creating a choice to explore one of two neighboring arms, as would be the case in each of the subsequent tactile discrimination phase. Crucially however, during both phases of habituation, the surface of the radial maze was entirely covered in the surface type (smooth or irregular) to which a given animal was to be assigned during the subsequent R1 tactile discrimination training. In this way, even during habituation, animals assigned to learn to associate, for example, the irregular surface with reward location had no prior experience the smooth surface being associated with reward, and vice versa.

Tactile discrimination: R1 tactile discrimination training (figure 1) was conducted either under low red lighting, in which case the radial maze was also surrounded by a black curtain to fully conceal any visual extra-maze spatial cues, or, when the option became available to us, under infra-red light (i.e. total visible darkness, precluding the need for curtain surround). No difference in learning rates or other behaviors was remarked between the low red light with curtain configuration versus the infra-red light configuration. The major advantage of the latter was ease of set-up and physical access for the experimenter to the radial maze. The rationale behind conducting this task in darkness was to deprive the animal of its capacity for visual spatial orientation, thereby obliging it to rely on other sensory inputs, notably tactile. Ethanol (70%) was also used to give the radial maze a particular odor-based context. Both of these sensory

particularities, visual and olfactory, would also allow us to maximize the novelty of the second environment in the second phase of our project, covered in the follow-up paper to this one (Stevens et al., 2022b).

4 of the 8 arms, plus their corresponding central platform segments, were covered with smooth surface panels, the remaining 4 arms and platform segments with irregular surface ones. Each animal would learn that only one of the two surfaces was predictive of reward location (S1). The configuration of the two surface types could be modified in various ways from one session to the next. At the beginning of the first session only, once the animal had been placed in the central platform, 40 seconds were allowed before the experimenter launched the task. This interval was intended to let the animal make an initial exploration of the central platform and become aware of the novelty of the environment; i.e. no longer composed of just one, but rather two distinct surface types. Once these 40 seconds had passed, the training session was launched. Each trial of R1 training began with the opening of a contiguous pair of doors, giving access to two contiguous arms; one smooth surfaced arm and one irregular surfaced arm. The animal then had to choose which of these arms to visit, having already learned from the habituation phase that a food reward could be found in the distributor at the distal end of the arms. Once the animal was detected in the most distal zone of one of the accessible arms, the door to the unchosen arm closed automatically, denying the animal the opportunity to further revise its choice. The implication here being that, prior to that point, the animal could, in fact, revise an initial choice by retracing its steps and choosing the other arm instead. In order for the trial to end, the animal had to enter the most distal section of one arm (where, if it had chosen the S1 arm, it found a food reward), and then return to the central platform, whereupon the door of the arm just visited would close behind it. Subsequently, following an inter-trial interval of 5 seconds and once the animal was detected in a zone of the central platform opposite to them, the next pair of doors in the pre-programmed sequence opened. Across trials, the relative left and right position of S1 and S0 was counter-balanced. Training consisted of either one or two sessions per day, with each session composed of between 16 and 36 trials (all session sequences open and available, see below). Animals were exposed to a combination of arm pair presentations designed to expedite their learning of the tactile discrimination rule. In early sessions, sequences were composed of a combination of repeated consecutive presentations of a same pair plus pseudo-randomized presentations

of all available pairs. The aim of the repeated sequences was to explicitly lead the animal to inhibit its innate drive to alternate. As the animals approached criterion level across sessions, the arm pair presentation sequences became progressively pseudo-random, as this was ultimately the only reliable test that responses were based solely on tactile discrimination and not on other sources, such as body turn memory, etc. The final sessions of tactile discrimination training were therefore fully pseudo-random trial sequences. Performance criterion was fixed as follows: animals had to attain either an average of at least 75% correct responses across two of the final pseudo-random sessions, or above 70% if they had performed above 80% on at least one session. In experiments with a control group, control animals were rewarded on every trial regardless of which surface, S1 or S0, they chose. All young C57Bl6/J mice in the experimental group were trained until they reached criterion. In the case of the D_1-CB_1-KO and aged mice experiments presented here, not all animals reached criterion (see Results).

Analysis

All raw data extraction, analysis, statistical comparison, and graphical representation was generated using custom codes written in Python (Van Rossum & Drake, 2009) thanks to the pandas (Reback et al., 2020), numpy, pingouin (Vallat, 2018), bioinfokit (Bedre, 2021), matplotlib (Hunter, 2007), and seaborn (Waskom, 2021) libraries. All code is open source and available at https://github.com/metaphysiology. Here we give brief details about the behavioral parameters we analyzed.

Decision latency: The time taken by each animal between the instant when a trial began (doors of the current trial pair open) and the instant when the threshold between the central platform and the arm of the animal's definitive choice was first crossed (decision latency, milliseconds). We were further able to classify these decision latencies according to various factors such as surface type of the definitively chosen arm, etc.

Run time: The time taken for animals to travel the distance from the threshold of the definitively chosen arm to the reward-distributor containing distal extremity (run time, milliseconds). As above, we could then classify this measure according to whether the definitive choice was S1 or S0, etc.

Choice revision: During certain trials, animals crossed the threshold into one arm of a pair but, prior to entering its most distal zone (which would trigger the closing of the door to the unchosen arm), revised their choice by exiting the arm again. At this point animals could either choose to explore the other arm of the pair or (more rarely) re-enter the same arm. As long as the distal zone of either arm had not been entered, this process could technically continue indefinitely. We developed a novel analysis to quantify this behavior, which we took to be an occasional external and physical manifestation of the ongoing cognitive decision-making process. On a given trial, each additional crossing of either of the two central platform-to-arm thresholds, in the direction from the platform towards the arm only, was quantified as one choice revision. Each choice revision was quantified as a 'KOOK' unit, capturing the fact that some choice revisions were ultimately error-inducing, 'KO', while others were rectifying, 'OK'. In practice, when choice revision occurred, the mean number of KOOKs in a single trial was 1.11 and the median 1, but KOOK values sometimes went far higher, with 10 being the highest number of choice revisions we observed in a single trial. Hence, it must be noted that this physically manifest choice-revision behavior was subject to high inter- and intra-individual variability, with some animals having a higher tendency to manifest it than others, independently of any other relevant factors such as performance, etc., (supplementary figure S.2c, the cumulative quantification of choice revisions reveals the width of the distribution of this behavior). This choice revision behavior is similar but not identical to what is described in the literature as vicarious trial-and-error (VTE) (Redish, 2016). (See boxes on run time and choice revision in Discussion.)

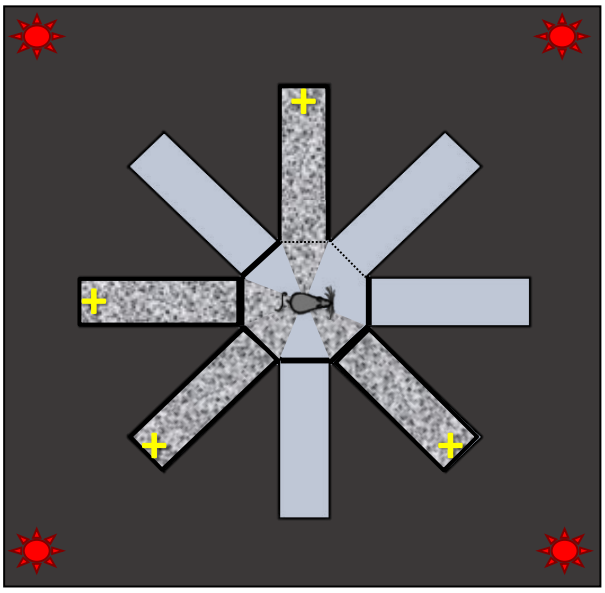


Figure 1 – "Indoctrination" learning radial maze environment.
Experiments were conducted in darkness under low-red or infra-red lighting. Solid grey represents smooth surface, dappled grey the irregular surface. Yellow '+' symbols represent location of food reward. Which surface was rewarded (S1) was counter-balanced between animals.

Results

1. Acquisition and expression of a binary choice-based tactile discrimination foraging rule.

We trained mice under a tactile discrimination-based reward-location association rule (R1) to choose, trial after trial, between the two contiguous radial maze arms of a sequence of arm-pairs, each arm of which was covered with one of two distinct surface types, one predictive of reward location (S1), one predictive of absence of reward (S0). This training was conducted in conditions of zero or almost zero visibility, i.e. without any extra-maze spatial cues, thereby constituting a classical stimulus-response (S-R) task where the stimulus in question was tactile (McDonald & Hong, 2004).

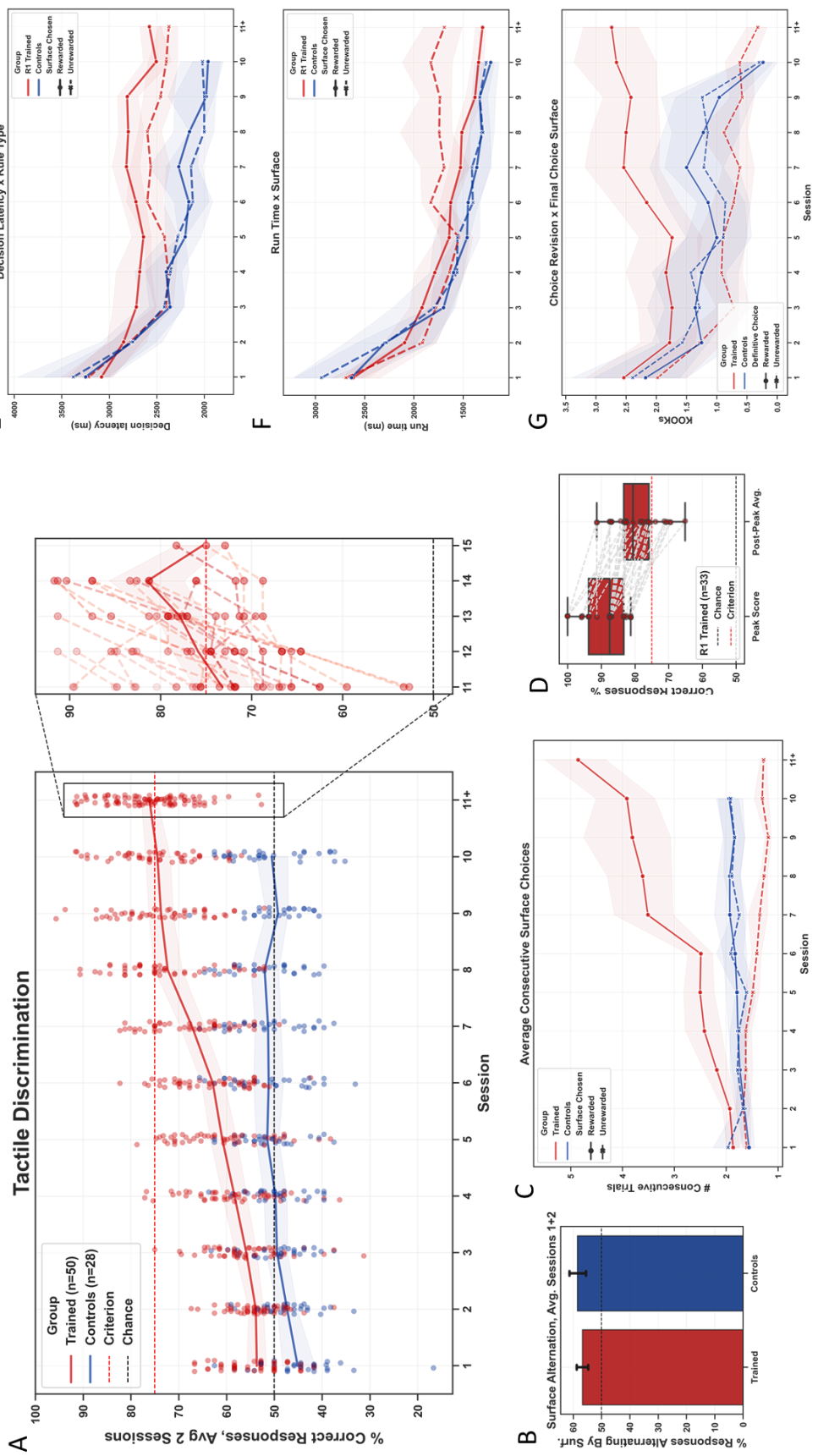

A
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Session
Group
Trained (n=50)
Controls (n=28)
Criterion
Chance
B
Surface Alternation, Avg. Sessions 1+2
% Responses Alternating By Surf.
Trained
Controls
C
Average Consecutive Surface Choices
Consecutive Trials
Session
Group
Trained
Controls
Surface Chosen
Rewarded
Unrewarded
D
Correct Responses %
R1 Trained (n=33)
Chance
Criterion
Peak Score
Post-Peak Avg.
E
Decision Latency x Rule Type
Decision latency (ms)
Group
R1 Trained
Controls
Surface Chosen
Rewarded
Unrewarded
F
Run Time x Surface
Run time (ms)
G
Choice Revision x Final Choice Surface
KOOKs
Session
Definitive Choice

Figure 2 – Behavioral responses to repeated training on a tactile win-stay stimulus-response rule.

Experimental animals are represented in red, controls in blue. All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. (**A**) % correct tactile discrimination responses over repeated training sessions, displayed as rolling averages over 2 sessions (as per definition of criterion level for end of training: 75% correct responses averaged over two sessions). R1 trained animals (n=50) represented in red, control animals (n=28) represented in blue. Controls were rewarded on all trials, regardless of surface chosen. Curves represent mean population score, dots represent individual performances. Zoombox shows additional training sessions needed for certain experimental animals to reach criterion. (**B**) Initial surface alternation behavior averaged across first two sessions in both experimental and control populations. Both populations alternated according to surface in their trial choices significantly more than chance level during these sessions; error bars = s.e.m. (**C**) Average number of consecutive S1 versus S0 choices per session. Space between error bands shows R1 trained population began making consecutive S1 choices significantly more than S0 choices beginning from session 3, a behavior never seen in controls. (**D**) Individual peak to average post-peak performance comparison for those R1 trained animals (33/50) who had highest performance prior to final training session. Drop in performance post-peak highly significant, with 7/33 animals falling back below criterion R1 level of 75%. (**E**) Median decision latencies by population and by surface. Decision latency significantly higher in R1 trained animals compared to controls, and significantly higher within group between S1 and S0. (**F**) Median run times (post-definitive choice) from door threshold to distal zone of arm by population and by surface. Controls, rewarded on every trial, rapidly develop faster run times than R1 trained population who, with repeated training, develop particularly long run times specifically on unrewarded S0 arms, but are also slower on always rewarded S1 arms than controls. (**G**) Mean total choice revision (animal initially crosses threshold into one arm but revises its choice before reaching distal zone) by population and by surface. R1 trained population began to engage in significantly more choice revision behavior than controls as early as session 3, with revisions increasingly and significantly favoring S1 final choices rather than S0 final choices.

1.1 Mice are innately sensitive to tactile differences and can use them to navigate a radial maze.

Tactile discrimination behavior: With extensive training, all young C57Bl6/J experimental mice reached the criterion level of 75% correct R1 responses averaged across two consecutive sessions (mean population performance in final session 83%, n=50, combined results from four iterations of the experiment; figure 2a, red line = population mean, red dots = individual performances). Animals in these experiments were trained until this criterion was reached, resulting in a differential total of training trials/sessions per individual (figure 2a zoom-box, mean trials to criterion = 265, spread across 11 to 15 sessions; T-test with Welch correction between overall mean and chance level of 50%, $t(595) = 31.1$, $p < 0.0001$). Control mice, rewarded on every trial during training sessions regardless of which surface they chose, developed no preference for either surface (control mice, n=28; figure 2a, blue line and dots; control mice performed between 7 and 10 training sessions, mean total trials = 222; T-test with Welch correction between overall mean and chance level of 50%, $t(277) = -0.62$, $p = 0.533$). In one of the experiments pooled into the analysis above (supplementary figure S.1i), we tested one group of mice on a within-maze tactile configuration which remained fixed in each session ('Fixed', n = 10), while for another group we altered the within-maze tactile configuration on a session-by-session basis ('Fluid', n = 10). The rationale behind this experiment was to test whether mice in the 'Fixed' group would, over time, form and furthermore make use of a tactile-based cognitive map of the radial maze. As can be seen by the 95% CI error bands in the figure, however, this environmental difference had no impact on R1 performance, adding evidence to the idea that R1 responding was primarily striatal rather than hippocampal. In short, whether or not mice in the 'Fixed' group did form a tactile cognitive map of the radial maze, it neither aided nor impeded them in reaching criterion performance in the S-R task.

Tactile spontaneous alternation: During early exposures to the environment (sessions 1 and 2), both the experimental and control groups displayed a robust, significantly greater than chance level tendency to alternate their arm choices from one trial to the next according to surface type (figure 2b; T-test, experimental group $t(99) = 6.5$, $p < 0.0001$, controls, $t(55) = 5.35$, $p < 0.0001$). In the control group only, reinforced on every trial regardless of surface choice, this tendency remained significantly above chance level until the fifth session (supplementary figure S.1a; T-test, $p < 0.05$ per session

up to session 5; trend continued but $p > 0.05$ in sessions 6 to 10). Individual levels of initial surface alternation, a potential proxy for measuring strength of innate tactile-based exploratory drive, did not, however, correlate with subsequent global R1 performance in the experimental mouse group (supplementary figure S.1b, $R^2 = 0.01$). Thus, initially, each of the two surfaces was explored on average on no more than two consecutive trials before animals alternated surface. While this was observed in the experimental group in the first two sessions, the average number of consecutive trial choices towards S1 began to be significantly higher than those towards S0 beginning as early as the third session, indicating at least some level of S1-reward association acquisition by that stage (figure 2c; S1 = unbroken red lines in figures; S0 = dashed red lines in figures; pairwise t-tests with Bonferroni adjustment; session 3 $t(98) = 4.16$, $p = 0.0007$; all subsequent sessions, $p < 0.0007$; all sessions overall difference $t(644) = 18.7$, $p < 0.0001$). In control animals, however, regular surface alternation persisted across all sessions with no significant difference between the average number of consecutive responses on either surface in any session (figure 2c, blue lines in figures; pairwise t-tests with Bonferroni correction $t(554) = 0.24$, $p > 0.9$). The consistently low average number of consecutive trials towards S0 in experimental mice demonstrated very little variance across individuals as compared to the highly variant average number of consecutive trials towards S1 (see 95% CI error bands, figure 2c). Since initial strength of exploration was not predictive of performance, variance in the average number of consecutive S1 choices in the experimental group may be putatively due to variance in the strength of individual inhibition of the exploratory strategy.

Persistent R1-antagonistic exploratory behavior: Even when the average number of consecutive trial choices towards S1 reached its peak (figure 2c; mean consecutive correct choices, 4.9), this number corresponded to less than a quarter of the total number of trials comprising the session in question (23 or 24 trials). Even the average *maximum* streak of consecutive correct choices (mean of max correct choices, 9.8) corresponded to less than half the total number of trials in the corresponding session (supplementary figure S.1c-e). Around one third of animals from the experimental group (17 out of 50) attained their peak performance during their final session of training. In the other 33 animals, we were able to observe a significant drop of 'very large' effect size on the Cohen scale when comparing individual peak performance score to individual averaged post-peak performance score, with 7 of the animals' performance averages dropping

back below the criterion level of 75% post-peak (figure 2d, paired t-test between peak and post-peak performances, $t(32) = 8.33$, $p < 0.001$, unbiased Cohen effect size, $d = 1.44$). We also noted after how many trials in each session, on average (mean of discrete value), the experimental group first explored an unrewarded S0 arm (i.e. per session initial exploratory choice). We found that animals from the trained population did not, on average, first explore S0 significantly later than animals from the control group until session eight (supplementary figure S.1f). Even in the final sessions, however, the timing of R1 trained animals' initial S0 choices occurred earlier during sessions than expected. To investigate this, we modelled a predicted distribution of initial S0 choices using the R1 performance values from each individual animal's penultimate training session, then compared this to the actual S0 choice data from the final session (supplementary figure S.1g; trained group data = red bars, control group data = blue bars, trained group predictions = orange bars; predicted distribution represents mean values of 1000 modelled iterations). While the control group chose between S1 and S0 almost perfectly randomly, in accordance with their lack of surface preference (i.e. ~50% first chose S0 on the first trial, ~50% of the remainder on the second trial, etc.), the tactile choice behavior of trained animals in initial trials did not correspond to their R1 performances from the previous session. Notably, while the predicted distribution showed no animals choosing S0 on the first trial, the data showed that almost 20% of trained mice in fact did so. Overall, trained mice were nearly 3 times more likely to choose S0 within the first three trials of the final R1 session than our model based on their R1 performances from the previous session indicated. Correspondingly, the timing of the first (and in some cases only) S0 choice also did not correlate with individual R1 performance during the final session, indicating that initial exploratory behavior was not a direct function of strength of overall R1 expression in experimental animals (supplementary figure S.1g + h, $R^2 = 0.03$). This corresponds with the possibility that sustained R1 expression relies more on the strength of active and ongoing inhibition of the exploratory drive than it does on strength of R1 acquisition, and this could explain why we see most exploration towards the beginning of post-learning sessions, putatively prior to inhibition of exploration being engaged upon initial introduction into the environment. Furthermore, as we saw above, strength of initial innate exploratory drive in sessions 1 and 2 also did not predict R1 performance, again fitting the putative scenario in which it is the strength

of the cognitive capacity to inhibit spontaneous exploration that is the primary contributor to R1 expression.

1.2 Cognitive behavioral analysis of decision-making and execution behavior confirms decoupling of R1 acquisition and expression.

Beyond trial-by-trial surface choice behavior, the experimental group was also distinguishable from the control group by significant differences in fine-grained in-trial cognitive behaviors.

Decision latency: Following an initial global decrease in both groups between the first and second sessions, the decision latency (i.e. time elapsed from start of trial to instant of crossing threshold of definitively chosen arm) only in the experimental group began to level off from session 3 onwards, with values on S1 choice trials consistently higher within group compared to S0 choice trials, while decision latency in the control group, independently of chosen surface, continued to decrease steadily until the end of the R1 training phase (figure 2e; curves represent the per session population median of the individual per session median decision time values, grouped according to surface of the definitive arm choice; 'Group' and 'Surface' differences; two-way ANOVA with pairwise Tukey HSD post-hoc test; significant effect of 'Group', $F(1, 1741) = 21.5$, $p = 0.001$, with significant effect of interaction 'Group*Surface', $F(1, 1741) = 5.7$, $p = 0.017$; repeated measures ANOVA revealed a significant effect of 'Surface' in the experimental group, $F(1, 49) = 9.43$, $p = 0.003$, but not in control group, $F(1, 27) = 1.5$, $p = 0.23$; 'Group' and 'Session'/repeated training differences; repeated measures ANOVA with Greenhouse-Geisser correction within 'Session' for each group, from session 3 to end; experimental group, $F(8, 392) = 2.19$, $p = 0.058$; control group, $F(7, 189) = 4.15$, $p = 0.036$. The choice of session 3 here is not arbitrary but corresponds to the session where the experimental group no longer significantly alternated by surface and instead began to increase their average consecutive number of S1 choices, see supplementary figure S.1a and figure 2c). In short, decision latency was globally higher in experimental animals, who had to exploit only part rather than explore all of the radial maze in order to obtain food rewards. Strikingly, decision latency was also specifically higher within the experimental group on trials where the 'exploit' surface (S1) rather than the de facto 'explore' surface (S0) was chosen, seeming to reveal a hierarchy of cognitive effort in the

inhibition of spontaneous behaviors, being a supplementary inhibitory cognitive effort the control group did not need to engage in.

Post-choice run time: Over the course of the first five sessions, global run time (i.e. time elapsed between instant of crossing threshold of definitively chosen arm and instant of entering the distal, reward distributor containing zone of that arm) decreased significantly, independently of surface-arm choice and at a comparable rate in both the experimental and control groups (figure 2f; curves represent the per session population median of the individual median run times; repeated measures ANOVA within 'Session', $F(10, 780) = 75.7, p < 0.0001$). Beginning in session 3, however, there emerged a reliable trend for overall, surface independent run times to be lower in the control group compared to the experimental group. ANOVA tests revealed this trend to be statistically significant when averaged across training sessions (one-way ANOVA with Tukey HSD post-hoc, significant effect of 'Group', $F(1, 1743) = 6, p = 0.014$). Finally, beginning as a trend in session 6 and becoming more significant over subsequent training sessions, run time increased on trials where animals from the experimental group, but not the control group, chose S0 arms, while it simultaneously decreased when they chose S1 arms (figure 2f; repeated measures ANOVA within experimental group, significant effect of 'Surface', $F(1, 49) = 48.8, p < 0.0001$; post-hoc pairwise t-tests revealed significance between surface run times began at session 7, $t(98) = 2.95, p = 0.039$). Thus, control animals, always rewarded regardless of choice and thus putatively having greater post-choice confidence, displayed globally lower run times than experimental animals, who in turn had significantly lower run times on S1, the always rewarded surface, than on S0, the never rewarded surface. This demonstrates a clear post-choice cognitive differentiation between the two surfaces reflecting the strength of acquisition of the S1-reward association.

Choice revision: Overall, the experimental group engaged in significantly more choice revision behavior (i.e. initially entering one arm of the trial pair but revising that choice prior to reaching the distal zone) than the control group (figure 2g, one-way ANOVA with Tukey HSD post-hoc, significant effect of 'Group', $F(1, 1746) = 12.9, p = 0.001$). A significant trend only in mice from the experimental group to revise their choice in a rectifying (i.e. terminating on S1 arms) rather than error-inducing (i.e. terminating on S0 arms) manner was seen to emerge early in training, another indication of successful

R1 acquisition (repeated measures ANOVA, significant effect of 'Surface', $F(1, 49) = 203$, $p < 0.0001$; pairwise t-tests reveal first significant effect of 'Surface' occurred in session 3, $t(76) = 3.45$, $p = 0.0009$). Looking at the performance from the final session, we also searched for indications that a process of proceduralization of R1 behavior in the experimental group may have given rise to decreasing choice revision behavior, as predicted by (Redish, 2016). Firstly, however, mean population levels of choice revision did not decrease as a function of training. Secondly, no relationship emerged between level of choice revision behavior and either strength of R1 performance or S1 run time (supplementary figure S.2a-b, $R^2 = 0.00$, $R^2 = 0.026$, respectively). Neither of these findings can be taken to be conclusive, however, since, with sufficient over-training, choice revision may effectively eventually decrease, in line with predictions from the literature. Nevertheless, it is difficult to predict just how much training that would require in our particular environmental conditions.

Overall, significant between- and within-group differences emerged as a function of training: higher overall decision latency in experimental group, primarily driven by within-group S1 choice decision times; higher overall run time in experimental group, amplified by increased within-group values on S0 arms; more overall choice revision in experimental group, driven by marked within-group trend to revise towards S1 more than towards S0. Each of these findings demonstrates that imposing an exploitatory response behavior in an environment where exploration is the spontaneous response gives rise to a measurable and significant increase in cognitive effort, which we suggest is primarily the effort of having to actively inhibit the exploratory drive *in order to* express R1.

2. Independence of R1 acquisition and expression investigated via manipulation of striatal function.

As we have just seen (figure 2a, f, g), detailed behavioral analysis revealed that mice from the experimental group demonstrated cognitive signs of having acquired the S1-reward association up to 10 sessions prior to reaching criterion R1 performance level. If successful exploitation of R1 did in fact require inhibition of an innate exploratory drive, i.e. the very basis of spontaneous alternation memory models in rodents, we hypothesized that striatal, specifically direct and indirect pathway, functions such as

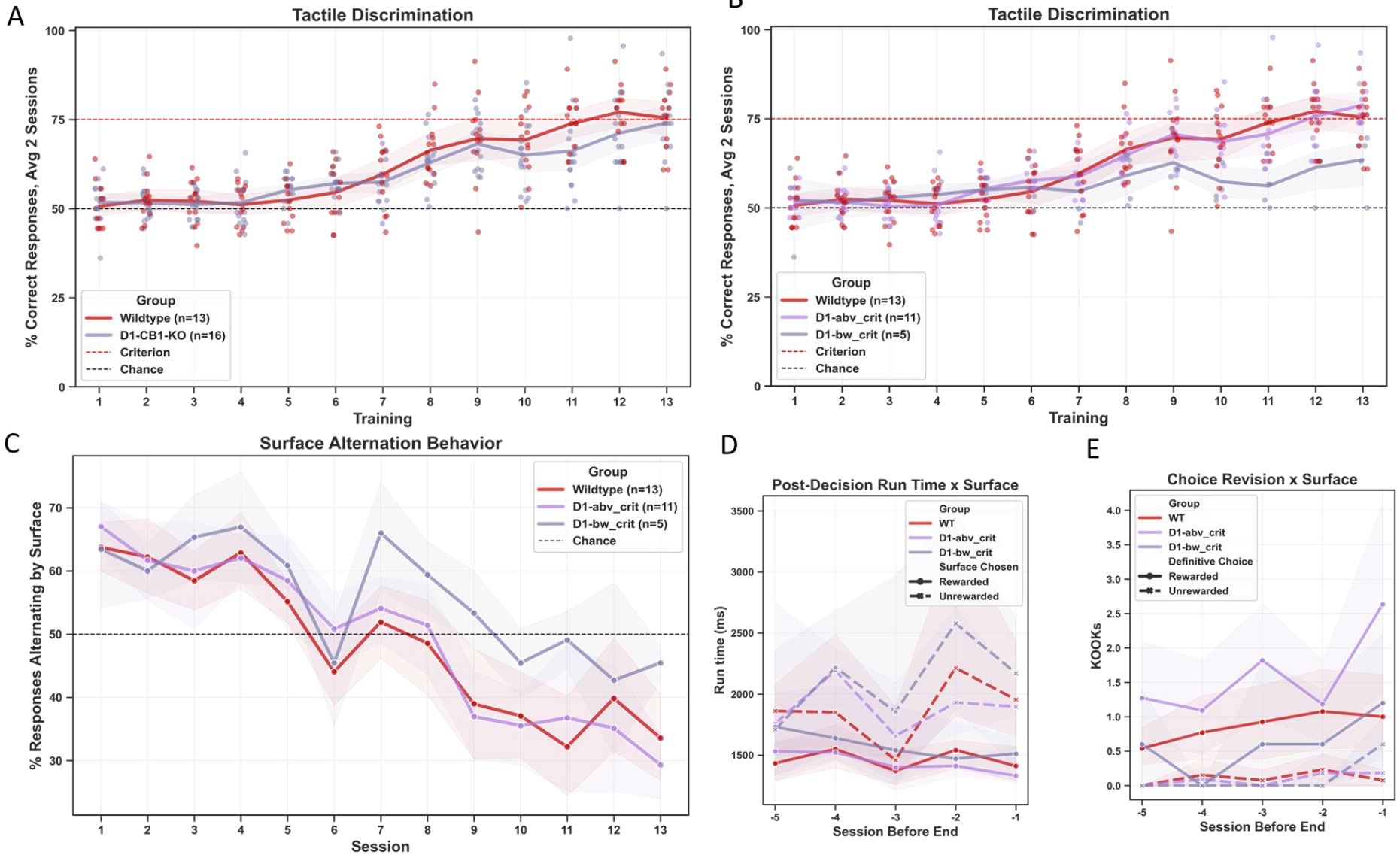

A
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Training
Group
Wildtype (n=13)
D1-CB1-KO (n=16)
Criterion
Chance
B
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Training
Group
Wildtype (n=13)
D1-abv_crit (n=11)
D1-bw_crit (n=5)
Criterion
Chance
C
Surface Alternation Behavior
% Responses Alternating by Surface
Session
Group
Wildtype (n=13)
D1-abv_crit (n=11)
D1-bw_crit (n=5)
Chance
D
Post-Decision Run Time x Surface
Run time (ms)
Session Before End
Group
WT
D1-abv_crit
D1-bw_crit
Surface Chosen
Rewarded
Unrewarded
E
Choice Revision x Surface
KOOKs
Session Before End
Group
WT
D1-abv_crit
D1-bw_crit
Definitive Choice
Rewarded
Unrewarded

Figure 3 – Deletion of CB1 receptors from D1-positive neurons impacts expression but not acquisition of R1.

All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. (**A**) Evolution of % correct tactile discrimination responses over repeated training sessions, displayed as rolling averages over 2 sessions (as per definition of criterion level: 75% correct responses averaged over two sessions). Wildtype animals (n=13) represented in red, D_1-CB_1-KO (n=16) in blue-gray. Curves represent mean population score, dots represent individual performances. The D_1-CB_1-KO population dropped slightly but significantly below wildtype littermates in expression of R1. (**B**) D_1-CB_1-KO animals were then subdivided into two populations: those who did reach criterion, D_1-CB_1-KO_{abv_crit} (n=11) represented in lilac; and those who did not, D_1-CB_1-KO_{bw_crit} (n=5) still in blue-gray. This subdivision of the D_1-CB_1-KO population was intended to reveal where above and below criterion animals did and did not differ in parameters other than R1 expression. (**C**) D_1-CB_1-KO_{bw_crit} displayed a rebound in exploratory surface alternation behavior which corresponded to the same point (session 7) where R1 expression in the D_1-CB_1-KO_{abv_crit} and wildtype populations began to rise above theirs. Overall, the D_1-CB_1-KO_{bw_crit} group alternated by surface significantly more than both the D_1-CB_1-KO_{abv_crit} and wildtype groups. (**D**) Median run times by group and by surface. All three groups displayed significantly higher run times on the unrewarded S0 surface compared to the S1 surface, a cognitive indicator of S1-reward association acquisition. No significant difference in S0 run times between groups indicated equal levels of R1 acquisition. (**E**) Mean total choice revision. D_1-CB_1-KO_{bw_crit} did perform significantly less choice revision towards S1 than D_1-CB_1-KO_{abv_crit}, indicating that capacity for choice revision may be a significant contributing factor in R1 expression.

novel action "chunking" and action selection (Graybiel, 1998) would be central to resistance towards expression of R1 and its subsequent exploitation. As a complement to this hypothesis, we also reasoned that these same specific striatal functions would not impact acquisition of the S1-reward association rule itself, since the strict acquisition implied no conflict with prior cognitive contents or tendencies. In order to explore this scenario, we used a genetic approach, employing the D_1-CB_1-KO transgenic mouse line to potentiate the net inhibitory signal of the D_1-expressing medium spiny neurons of the direct pathway. Our hypothesis was that, in turn, selection of innate exploratory responses would also be potentiated in D_1-CB_1-KO animals, potentially disrupting a gradual, indirect pathway-mediated inhibitory no-reward coding "shift" of strategy towards the tactile S-R action required for high R1 performance. To enable direct comparison between all experimental animals, R1 training sessions in the following experiments were capped at 13 for all mice, with no additional training sessions for those who did not reach criterion by that stage. The final 5 sessions, comprising 131 trials in total, were composed entirely of pseudo-randomized trial-pair sequences.

2.1 Deletion of CB1 from D1 positive neurons negatively impacts expression, but not acquisition, of an exploration-antagonistic S-R rule.

We observed that, averaged across the final 5 sessions of pseudo-randomized trial sequences, the wildtype group displayed significantly higher R1 expression than the D_1-CB_1-KO group (figure 3a, results pooled from two iterations of the experiment, both of which produced comparable results; WT, n = 13; D_1-CB_1-KO, n = 16; one-way ANOVA with pairwise Tukey HSD post-hoc, $F(1, 143) = 6.26$, $p = 0.013$). To investigate our hypothesis that any R1 exploitation deficit in D_1-CB_1-KO mice would be due to a reduced capacity for expression rather than acquisition of the S1-reward association rule, we divided the D_1-CB_1-KO population into two groups; those who did reach R1 criterion level within 13 training sessions and those that did not. The former group we labelled D_1-CB_1-KO_{abv_crit} (n = 11, including 3 borderline cases whose highest 2-session average R1 performance was 73.9% rather than 75%, but who had performed above 80% in at least one session), the latter group D_1-CB_1-KO_{bw_crit} (n = 5). The D_1-CB_1-KO_{abv_crit} group began to outperform the D_1-CB_1-KO_{bw_crit} group in terms of R1 expression as early as session 6 (figure 3b), giving rise to a significant overall difference

between the two groups in R1 performance averaged across the final 5 pseudo-randomized trial sequence sessions (one-way ANOVA with pairwise Tukey HSD post-hoc, $F(1, 78) = 39.07, p = 0.001$). Naturally, however, this difference in R1 performance must primarily be understood as an artefact of how the two groups were defined. More strikingly, when we analyzed the respective median run times between the above and below criterion groups, focusing again on the final 5 sessions, we found that both manifested significantly higher values on S0 than on S1 (figure 3d, one-way ANOVAs with Tukey HSD post-hoc, $D_1\text{-}CB_1\text{-}KO_{abv\_crit}$, $F(1, 107) = 36.1$, $p = 0.001$; $D_1\text{-}CB_1\text{-}KO_{bw\_crit}$, $F(1, 48) = 19.02$, $p = 0.001$; WT curve is displayed for visual comparative purposes but not further analyzed). Moreover, there was no significant difference between the groups in either overall run time or in S0 run time specifically (one-way ANOVA; overall run time between groups, $F(1, 157) = 0.27$, $p = 0.61$; S0 run time between groups, $F(1, 77) = 0.02$, $p = 0.88$). This difference in run time between S1 and S0 indicated that sensorial environmental feedback reached indistinguishable strength of cognitive "meaning" relative to R1 in both above and below criterion $D_1\text{-}CB_1\text{-}KO$ mice. With respect to choice revision behavior, we found that the $D_1\text{-}CB_1\text{-}KO_{abv\_crit}$ group did engage in significantly more choice revision behavior, specifically towards the rewarded surface, compared to the $D_1\text{-}CB_1\text{-}KO_{bw\_crit}$ group (figure 3e; two-way ANOVA on mean choice revision values revealed a significant 'Group*Surface' interaction, $F(1, 156) = 8.33$, $p = 0.007$; post-hoc pairwise Tukey HSD revealed key difference in '$D_1\text{-}CB_1\text{-}KO_{abv\_crit}$ S1' vs '$D_1\text{-}CB_1\text{-}KO_{bw\_crit}$ S1', $p = 0.001$). This suggested that R1 expression differences between the two groups could reside in a differential ability to actively inhibit (albeit sometimes after-the-fact via choice revision) the spontaneous and innate exploratory drive, i.e. precisely that putatively indirect pathway-mediated function we hypothesized would be relatively weaker in a scenario where direct pathway activity was potentiated. Interestingly, there was also a corresponding trend in 11 out of the 13 sessions for $D_1\text{-}CB_1\text{-}KO_{bw\_crit}$ animals to perform more surface-based alternation behavior, a behavior already linked with exploratory processes in our original experiments above, than the $D_1\text{-}CB_1\text{-}KO_{abv\_crit}$ or wildtype groups (figure 3c and figure 2b). Overall, $D_1\text{-}CB_1\text{-}KO_{bw\_crit}$ animals alternated by surface significantly more than the other two groups, another indication that the reason for their low R1 expression was related to exploratory drive and not weak R1 acquisition (one-way ANOVA with pairwise Tukey HSD post-hoc; $F(2, 374) = 5.1$; $D_1\text{-}CB_1\text{-}KO_{bw\_crit}$ vs $D_1\text{-}CB_1\text{-}KO_{abv\_crit}$, p

= 0.02; D_1-CB_1-KO_{bw_crit} vs wildtype, p = 0.006). Finally on this point, when the same animals were subsequently tested in an exploratory, alternation-based declarative memory task (Stevens et al. 2022b), D_1-CB_1-KO_{bw_crit} animals significantly outperformed D_1-CB_1-KO_{abv_crit} animals. This again strongly indicates that successful inhibition of the exploratory drive, rather than strength of acquisition of the S1-reward association per se, is the key factor in achieving strong R1 expression. Further complementary evidence for the fundamental role played in R1 expression by a certain equilibrium between the direct and indirect pathways also came from an experiment we ran in animals lacking CB_1 from *all* GABAergic neurons of the forebrain (therefore notably including both the D_1-positive cells of the direct pathway *and* the D_2-positive cells of the indirect pathway), which allowed us to posit that a new equilibrium had been reached between two "potentiated" neuronal populations. In contrast to the D_1-CB_1-KO population, this putatively striatally "balanced" transgenic line (named Dlx-CB_1-KO for the GABA-specific Dlx5/6 gene) displayed no observable difference in their R1 expression curve compared to their wildtype littermates (supplementary figure S.3a).

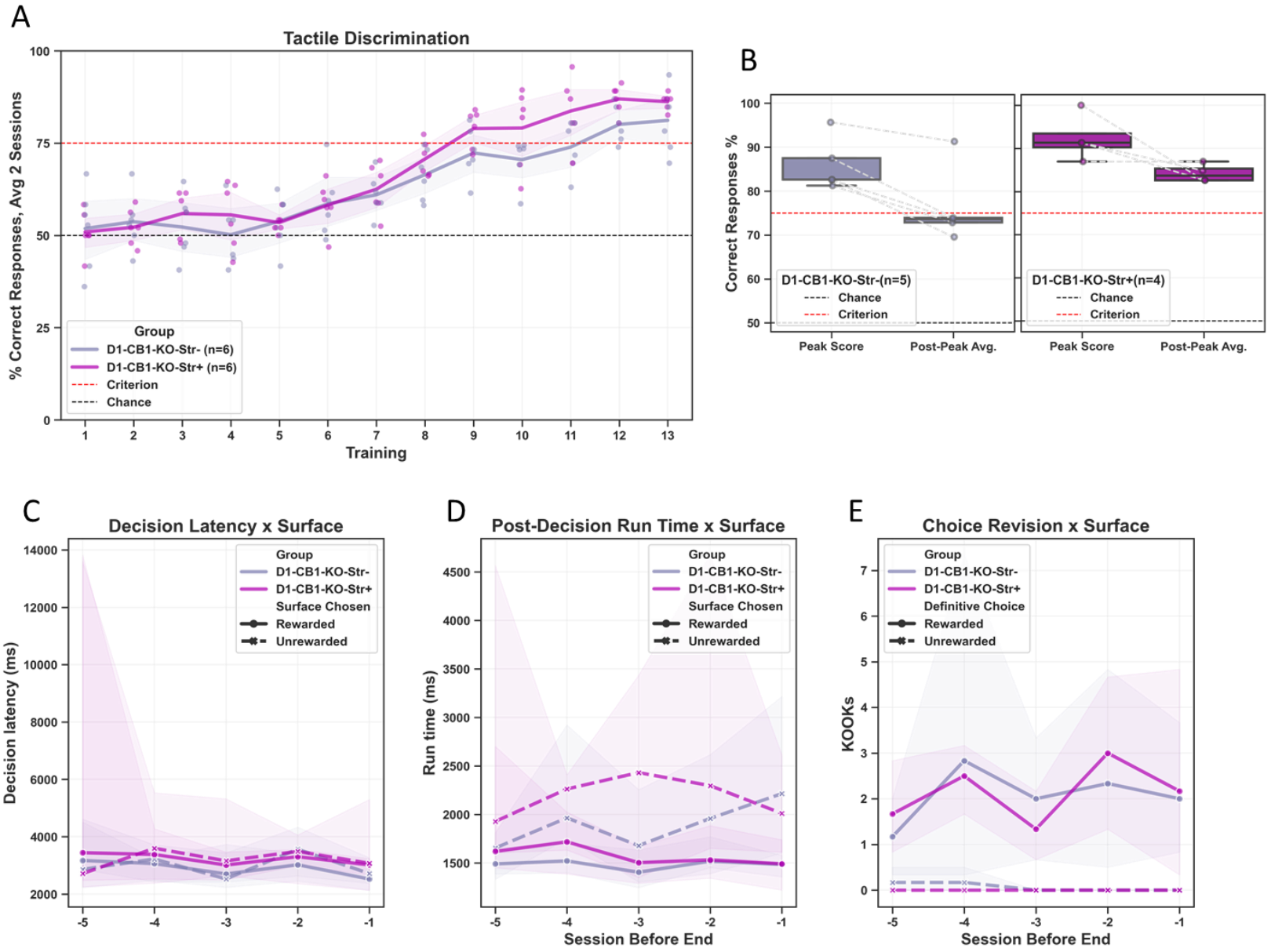

A
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Training
Group
D1-CB1-KO-Str- (n=6)
D1-CB1-KO-Str+ (n=6)
Criterion
Chance
B
Correct Responses %
D1-CB1-KO-Str-(n=5)
Chance
Criterion
D1-CB1-KO-Str+(n=4)
Chance
Criterion
Peak Score
Post-Peak Avg.
Peak Score
Post-Peak Avg.
C
Decision Latency x Surface
Decision latency (ms)
Group
D1-CB1-KO-Str-
D1-CB1-KO-Str+
Surface Chosen
Rewarded
Unrewarded
D
Post-Decision Run Time x Surface
Run time (ms)
Session Before End
E
Choice Revision x Surface
KOOKs
Definitive Choice
Session Before End

Figure 4 - Re-expression of CB1 receptors in D1-positive neurons recovers capacity for sustained R1 expression.

All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. (**A**) Evolution of % correct tactile discrimination responses over repeated training sessions, displayed as rolling averages over 2 sessions (as per definition of criterion level: 75% correct responses averaged over two sessions). D_1-CB_1-$KO_{Str-/-}$ (n=6) represented in blue-gray, D_1-CB_1-$KO_{Str+/+}$ (n=6) represented in magenta. Curves represent mean population score, dots represent individual performances. D_1-CB_1-$KO_{Str+/+}$ robustly overtook D_1-CB_1-$KO_{Str-/-}$ in R1 expression starting from session 7. (**B**) Peak to average post-peak performance comparison shows that D_1-CB_1-$KO_{Str+/+}$ not only expressed R1 more strongly but also more robustly than D_1-CB_1-$KO_{Str-/-}$, no individual falling below criterion following peak performance. This robustness of R1 exploitation we attribute to CB_1 re-expression giving rise to an over-expression, thereby boosting local inhibitory control in the direct pathway. (**C**) Median decision latencies by group and by surface. No significant differences were observed between the two groups in terms of decision latency. (**D**) Median run times by group and by surface. Similarly with run time, both groups displayed comparable differences between S0 and S1 run times, albeit with higher variability in D_1-CB_1-$KO_{Str+/+}$. (**E**) Mean total choice revision. Choice revision was also significantly biased towards final S1 choices, with no significant difference between the two groups.

2.2 Re-expression of CB_1 receptors in D_1 positive neurons of the direct pathway rescues performance of an exploration-antagonistic S-R rule.

In order to confirm that the deficit observed in overall R1 expression in D_1-CB_1-KO animals was indeed due to deletion of CB_1 from D_1 positive neurons of the direct pathway, we used a viral approach to re-express CB_1 receptors locally in D_1 positive neurons of the striatum (D_1-CB_1-$KO_{Str+/+}$, n = 6). If our general hypothesis regarding the contribution of inhibitory modulation of the direct pathway during R1 expression were accurate, then this targeted CB_1 re-expression should rescue wildtype levels of R1 performance compared to D_1-CB_1-KO animals injected with an empty vector virus (D_1-CB_1-$KO_{Str-/-}$, n = 6; see Materials & Methods). The behavioral results we obtained confirmed these predictions. Firstly, averaged across the final 5 pseudo-randomized trial sequence sessions, D_1-CB_1-$KO_{Str+/+}$ mice displayed significantly stronger expression of R1 than D_1-CB_1-$KO_{Str-/-}$ mice (figure 4a, one-way ANOVA with pairwise Tukey HSD post-hoc, $F(1, 58) = 14.33$, $p = 0.001$). Furthermore, we also observed less post-peak performance variance in the D_1-CB_1-$KO_{Str+/+}$ group compared to both the D_1-CB_1-$KO_{Str-/-}$ group and other groups from other experiments, visible in the fact that all individuals from the CB_1 reexpression group maintained above criterion expression even post-peak (figure 4b). Two other details are worth mentioning. Firstly, in this experiment all D_1-CB_1-$KO_{Str-/-}$ mice also reached criterion performance, in contrast to what we had observed in D_1-CB_1-KO mice. This relative improvement in performance could be related to the fact that these animals had undergone and recovered from surgery, or it could simply be the result of probabilistic variance which would average out to a lower performance if replicated with a larger population. Neither scenario, however, detracts from the significantly boosted performance of the D_1-CB_1-$KO_{Str+/+}$ group compared to the D_1-CB_1-$KO_{Str-/-}$ group. Secondly, since our targeted viral re-expression of CB_1 actually gives rise to an over-expression, compared to wildtype levels, of the receptor protein (both in per neuron absolute terms, due to the CAG promoter (Hitoshi et al., 1991), and in terms of expression in D_1-positive neurons which do not express CB_1 in wildtype animals), this apparent boosting, beyond mere rescue, of R1 expression lends strength to the idea that inhibitory modulatory control in the direct pathway is indeed central to successful inhibition of spontaneous, innate exploratory strategies. Interestingly, we also observed a significant trend for D_1-CB_1-$KO_{Str+/+}$ animals to have both higher overall decision latencies and higher overall run times compared to D_1-CB_1-$KO_{Str-/-}$ animals,

potentially indicating that the relative over-expression of CB_1 increased the amount of overall decision-making cognition via increased inhibitory processes in the direct pathway, in turn contributing to stronger R1 performance (figure 4c-d, one-way ANOVA with pairwise Tukey HSD post-hoc, final 5 sessions; decision latency, $F(1, 309) = 5.63$, $p = 0.018$; run time, $F(1, 117) = 4.56$, $p = 0.035$). Both groups also engaged in significantly more choice revision towards S1 than S0, but with no significant difference between the two groups (figure 4e). Though we may have expected to observe more manifest choice revision in the D_1-CB_1-$KO_{Str+/+}$ group, their increased decision latencies still indicate that they displayed more cogitation in resolving choice conflict than did the D_1-CB_1-$KO_{Str-/-}$ group.

2.3 Aged mice negatively impacted in expression but not acquisition of an exploration-antagonistic S-R rule.

As can be seen from the x-axis of figure 5a, R1 training of aged animals, as we had predicted, required more sessions than any of our previous experiments with young C57Bl6/J or transgenic animals. By session 13 (i.e. the number of training sessions fixed for the genetic approach experiments seen above) the mean R1 performance averaged across two sessions of the experimental population was only 63.9%, with only one individual having reached the 75% performance criterion for that session. This resistance to sustained expression of R1 was reflected in more persistent surface alternation behavior in aged compared to young experimental mice (compare, for example, sessions 7 in supplementary figures S.1a and S.4a). Consequently, we continued training with all animals, experimental and controls, for a further two sessions. By this point (session 15), 3 out of 11 aged mice from the experimental group still had not yet reached criterion levels.

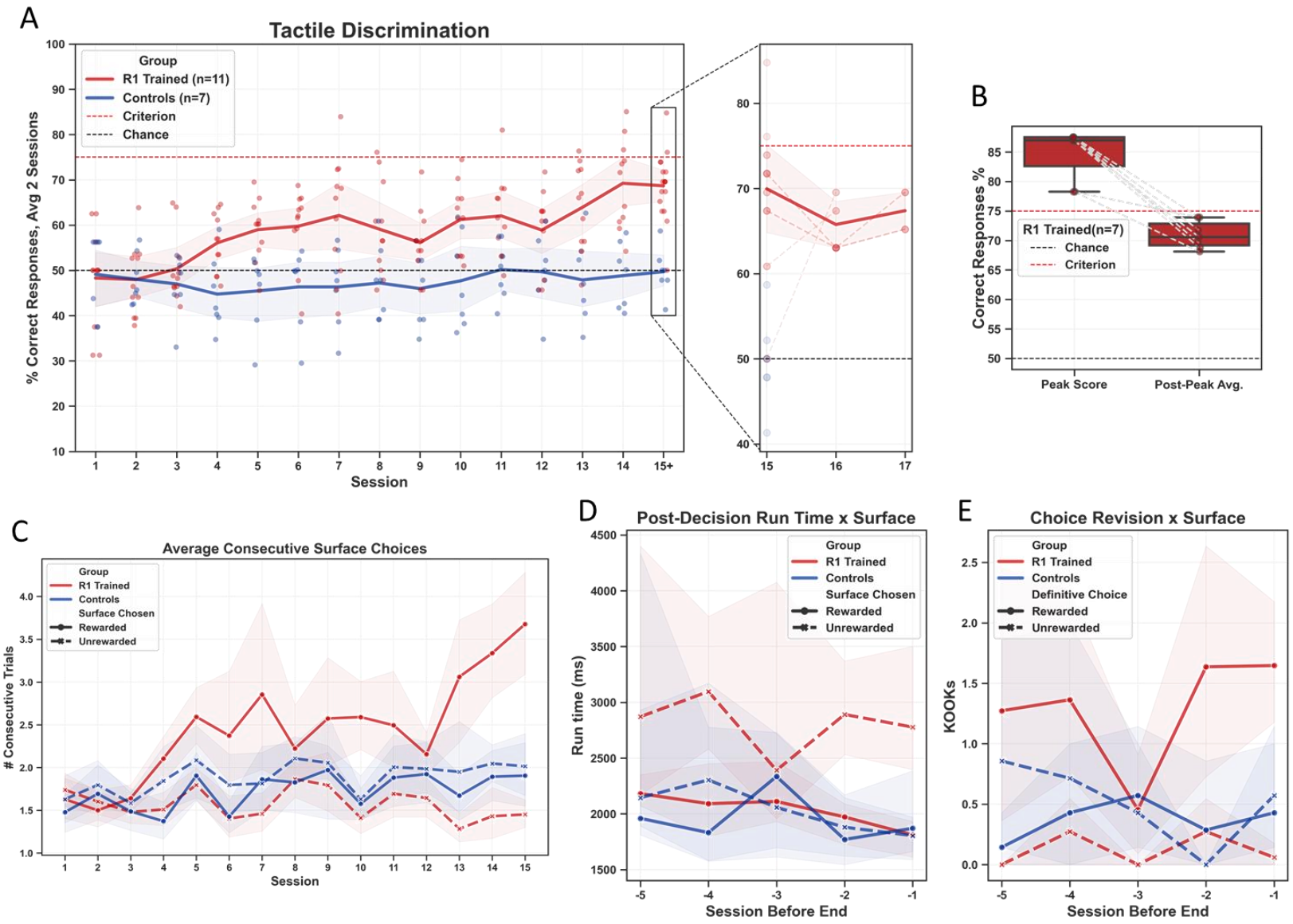

A
Tactile Discrimination
Group
R1 Trained (n=11)
Controls (n=7)
Criterion
Chance
% Correct Responses, Avg 2 Sessions
Session
B
Correct Responses %
R1 Trained(n=7)
Chance
Criterion
Peak Score
Post-Peak Avg.
C
Average Consecutive Surface Choices
Group
R1 Trained
Controls
Surface Chosen
Rewarded
Unrewarded
Consecutive Trials
Session
D
Post-Decision Run Time x Surface
Run time (ms)
Session Before End
E
Choice Revision x Surface
Definitive Choice
KOOKs
Session Before End

Figure 5 - Aged mice severely impaired in expression but not acquisition of R1.
All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. (**A**) Evolution of % correct tactile discrimination responses over repeated training sessions, displayed as rolling averages over 2 sessions (as per definition of criterion level: 75% correct responses averaged over two sessions). R1 trained aged mice (n=11) represented in red, aged controls (n=7) represented in blue. Controls were rewarded on all trials, regardless of surface choice. Curves represent mean population score, dots represent individual performances. Even after 15 sessions of R1 training, 3 aged mice had still not reached criterion performance. Additional training sessions (zoombox) were not sufficient to elicit stronger R1 expression in these animals. (**B**) Peak to average post-peak performance comparison revealed highly significant drop following peak performance, with all animals who reached peak performance prior to final session (n=7) subsequently dropping back below criterion level. (**C**) Average consecutive surface choices per session: R1 trained animals were consecutively choosing S1 arms significantly more than S0 arms already by session 4, indicating that the S1-reward association was acquired at this point. (**D**) Median post-decision run times by group and by surface. R1 trained animals had significantly higher run times on S0 compared to S1 across all 5 final pseudo-random R1 training sessions, again showing that R1 was robustly acquired. (**E**) Mean total choice revisions. Aged R1 trained animals significantly favored final S1 decisions when revising initial choice across all 5 final pseudo-random R1 training sessions.

Furthermore, out of those animals who had reached criterion, we observed a highly significant post-peak performance drop of 'huge' Cohen effect size in the 7 animals who reached criterion prior to their final session of training (figure 5b; paired t-test with unbiased Cohen effect size, $t(6)$ 7.3, $p = 0.0003$, $d = 3.9$). This provided still further evidence of the ongoing active inhibition of the exploratory drive necessary for sustained exploitation of R1. Difficulty not only in engaging but also in maintaining such a level of active inhibition could therefore be a fundamental dimension of age-related deficits in cognitive flexibility. Finally, it is also worth remarking that whereas surface alternation behavior in the young R1 trained population dropped significantly below 50% as early as session 7 (supplementary figure S.1a), in the aged population this did not happen until session 14 (supplementary figure S.4a), which is yet again strong indication that aged mice were more rigidly exploratory than young adult mice.

Naturally, we next asked the same question of our aged animals as we had of the D_1-CB_1-KO mice (see figure 3d), i.e. was their deficit in R1 exploitation due to an impairment in acquisition or expression of R1, if not both? Our initial hypothesis was the same as with the D_1-CB_1-KO mice; that any R1 performance deficit in aged mice would be primarily due to impairment of expression but not acquisition of the S1-reward association. Our observations confirmed this hypothesis. Firstly, although variance was much higher than in the results from young C57Bl6/J (due to both smaller population size and, precisely, to aged mice behaving more erratically with respect to exploiting R1), the greater number of average consecutive choices towards S1 compared to S0 approached significance as early as session 4 (figure 5c; pairwise t-tests with Bonferroni adjustment; session 4 $t(20) = 3.27$, $p = 0.058$). Having reached statistical significance in session 5, the magnitude of the difference in consecutive choices between S1 and S0, although always reliably present, subsequently fluctuated greatly from session to session (which we attribute to age-related decrease in the capacity to sustainably inhibit spontaneous exploration) until the final 3 sessions in which the difference was reliably highly significant (session 5, $t(20) = 3.91$, $p = 0.013$; sessions 13-15, least significant value, $t(20) = 4.37$, $p = 0.004$). The overall difference in S1 vs S0 average consecutive choices across all training sessions was also highly significant ($t(340) = 10.1$, $p < 0.0001$).

With respect now to the more fine-grained cognitive measures, in post-choice run time we observed a robust trend for experimental aged mice to have higher run times on S0

arms emerging as early as session 6, and this reached statistical significance across the final 5 sessions (figure 5d; one-way ANOVA with Tukey HSD post-hoc, $F(1, 120) = 48.18$, $p = 0.001$). Similarly, mean choice revision behavior in R1 trained aged mice also began to favor correction towards S1 beginning from session 6 and was significant across the final 5 sessions (figure 5e; one-way ANOVA with Tukey HSD post-hoc, $F(1, 120) = 41.18$, $p = 0.001$). Although in this experiment we did not test young mice simultaneously with aged mice, the run time results here correspond with those observed in our earlier experiments (compare figure 5d and figure 2f), with this difference that choice revision in young mice began to favor S1 earlier and was, globally speaking, more pronounced and less erratic than in aged mice (compare figure 5e and figure 2g; large differences in population size may also be an important factor here). Regarding decision latency, as with young mice, we observed globally higher decision times in trained animals than in controls, although this did not reach statistical significance across the final 5 pseudo-random trial sequence sessions (supplementary figure S.4b; one-way ANOVA between 'Group', $F(1, 190) = 0.21$, $p = 0.65$). That the difference between the two populations did not reach significance appears to have been due to especially high levels of variability in the control group rather than being due to a reduced phenotype in the experimental group. Finally, in order to doubly confirm our acquisition vs expression hypothesis, we also looked at the relative S1 vs S0 run times in only those three experimental animals who did *not* reach criterion R1 expression levels. Here again we saw the same robust trend, this time beginning in session 4, for S0 run times to be greater than S1 run times, indicating that the S1-reward association had indeed been acquired even by these particularly R1 expression resistant animals (supplementary figure S.4c). This provided important supplementary confirmation that the impairment in aged mice was indeed at the level of R1 expression rather than acquisition, and therefore that the key mechanism lacking vigor in older mice was inhibition of spontaneous exploratory strategies.

Discussion

The present study is the first to explicitly frame behavioral training of rodents that is intentionally antagonistic with respect to their spontaneous exploratory drive as a model of indoctrination, based on a broad definition of this concept as any mode of teaching which by its nature thwarts "open-mindedness" or natural curiosity (Callan & Arena, 2009; Taylor, 2017). The task itself can be understood as a simple state-action policy revision (Sutton & Barto, 2018), insofar as innate or "naïve" animal behaviors can be framed as evolutionarily preserved state-action policies, of which "exploration" is the one we expected rodents to spontaneously express in the experimental environment we had designed.

Innate behavioral phenotypes of complex organisms preserved across evolution, such as curiosity, are theorized to reflect those cognitive responses most likely to obtain an adaptive advantage across the range of environments with which a species has evolved (Lewontin & Levins, 2000; Levins & Lewontin, 1985; Lorenz, 1958; Schmalhausen, 1949). But what adaptive advantage could the potential to be indoctrinated carry? Novel, learned cognitive behaviors, acquired as a function of contingent particularities of a given environment in which the organism actually finds itself, can be either cooperative or competitive with respect to innate, unlearned behaviors (Reid et al., 2008; Staddon & Simmelhag, 1971). Impairments, not only in acquisition but also in expression of novel, environmentally contingent responses, risk suboptimal exploitation of environments or even, in some human and animal contexts, punishment. With respect to this, cognitive plasticity or flexibility itself has been conceptualized as a preserved feature of complex organisms, precisely enabling them to inhibit and go beyond the innate in acquiring and expressing novel behaviors in order to best exploit unpredictable environmental particularities (Wexler, 2011). Thus, although it may seem conceptually counter-intuitive, our results indicate that indoctrination, in the sense where this term implies the enforced suppression of natural tendencies towards exploration and curiosity, not only requires inhibitory cognitive flexibility but may be best understood as a coopting (Gould et al., 1979; Gould & Vrba, 1982), neuronal recycling (Dehaene & Cohen, 2007), or "hijacking" (Munro, 2015; Schultz, 2016) of it as a means of suppressing the spontaneous exploratory drive.

Tactile sensitivity and surface-based spatial alternation.

The aim of the task we have introduced here was to train mice to the point of significant exploitation of a tactile-based S-R rule, a rule specifically designed to be antagonistic towards their innate, spontaneous exploratory drive. The exploratory drive in question is precisely that which underpins the well-characterized and experimentally exploited rodent behavior of spontaneous spatial alternation (Lalonde, 2002; Olton & Samuelson, 1976; Richman et al., 1986). That extended training was needed for many animals to reach criterion R1 performance (i.e. 75% correct responses averaged across two sessions) invites a first objection that perhaps mice were not immediately sensitive to the surface distinction, or that mice who were slow to reach criterion were simply slow in acquiring the S1-reward association. With respect to the first objection, we had, on the contrary, predicted that since the task was conducted in the absence of visual spatial cues so mice would initially employ their sensory awareness of the distinct surfaces as a guide to direct exploration of the entire surface of the radial maze via a process of surface-based, as opposed to visual or proprioceptive, etc., spatial alternation. Such tactile alternation behavior would presuppose a capacity not only to sensorially discriminate between the two surfaces but also to cognitively contextualize their semantic relevance as spatial orienters (Colombo et al., 1990; McDonald & Hong, 2004). Accordingly, despite the fact that it was not reinforced, initial above chance level trial-to-trial surface-based alternation behavior is precisely what we observed in each iteration of our protocol, whether in young C57Bl6/J mice or in aged or transgenic animals. Moreover, control animals (rewarded on every trial regardless of whether they alternated or repeatedly chose one surface) maintained a trend for above chance level surface-based alternation throughout training (mean population score). The fact that all mice initially displayed above chance-level surface-based alternation, even though this behavior was not reinforced, also indicates that the exploratory drive in rodents is primitive with respect to classical visuo-spatial spontaneous alternation (Gaffan & Davies, 1981) and will recruit any relevant sensorial environmental cues afforded to an animal, including tactile, as a means of guiding exploration. Perhaps counter-intuitively, however, we found no evidence of any simple cognitive relation between strength of initial surface-based exploratory drive and subsequent capacity to inhibit it, as there was no correlation between individual initial surface alternation performance and subsequent strength of R1 expression. This was a first indication from our results that

the cognitive mechanisms required for inhibiting spontaneous exploration are to an important extent independent from the mechanisms underpinning both the strength of the innate exploratory drive and the strength of the acquired exploration antagonistic rule.

Exploitation requires active inhibition of exploration.

Demonstrating that resistance to exploitation was not the result of slow acquisition of the S1-reward association, our behavioral analyses in young C57Bl6/J mice revealed that this association could be detected in average consecutive surface choices, in decision latency, and in choice revision behavior as early as the third session, and in run time as early as the sixth session; i.e., long before sustained expression of R1 was reached. In other words, behavioral expression of the otherwise well acquired S1-reward association remained in persistent conflict with the more primitive drive to continue exploring the environment. Such decoupling of distinct memory expression systems in aged mice further corroborates earlier work on multiple memory systems from our team (Marighetto et al., 1999). This led us to conceive of "exploration" and "exploitation" *as they relate to the R1 environment* as distinct, integrated and unitary behavioral sequences, or "chunks" (Graybiel, 1998; Jin et al., 2014), putatively available for selection as competing strategies through the interactions of the direct and indirect pathways. In this regard, it was interesting to observe that in the experimental mouse group, median decision latency was significantly higher, not only overall when compared to control animals, but also specifically on trials where S1, the "exploitatory" option, was chosen as compared to trials where S0, the "exploratory" option, was chosen. This indicates that, in contrast to exploration (which is "spontaneous"), engaging in sustained exploitatory choices required an observable amount of additional cognitive effort. We suggest that this additional effort resides in a cognitive requirement to actively inhibit the innate exploratory strategy so that the acquired R1 strategy can be expressed.

Persistent conflict between the exploitation and exploration strategies was also observed in other parameters. When comparing peak performance to averaged post-peak performances in the majority of young C57Bl6/J who achieved their highest R1 performance *prior* to the final session (33 out of 50), R1 expression dropped sharply

and significantly following peak exploitation, with around 25% of animals dropping back below criterion levels post-peak. This again demonstrates that expression of R1 was not a simple matter of incrementally reinforcing or stamping-in the S1-reward association and related S-R response. Rather R1 expression also required active and ongoing inhibition of the more primitive exploratory drive. Results from the experiment where we added modulatory inhibitory control back to the direct pathway, via targeted viral reexpression of CB_1, further validated this interpretation: benefiting from increased inhibitory control (as described, targeted viral reexpression actually gives rise to an over-expression compared to baseline wildtype levels) animals were able not only to express R1 earlier and more reliably, but also to maintain a more even peak:post-peak performance ratio compared to control D_1-CB_1-KO animals injected with an empty viral vector. Conversely, in aged animals, with their well-characterized age-related deficit in cognitive flexibility generally and inhibitory control specifically (Coxon et al., 2012), R1 expression was slower to emerge, more erratic once it had emerged, and in the peak:post-peak performance ratio all aged animals concerned dropped back below criterion level, compared to just a subset of young mice. Taken together, these particular observations extend a somewhat neglected idea that reinforcement learning of one behavior may be as much, if not primarily, a question of the extinction, through non-reinforcement, of competing behaviors (Staddon & Simmelhag, 1971). In fact, by highlighting the role played by active inhibition, as opposed to mere extinction, of competing behaviors, our results also bring us closer to some of the most recent theories in the neuroeducation literature, notably the 3-system theory of the cognitive brain, in which the third system is precisely inhibitory control (Houdé, 2019).

Another manifestation of the same exploratory persistence was observed in the fact that, in the final session of R1 training, the experimental group first explored an S0 arm, on average, significantly earlier than could be predicted from a distribution of when initial S0 choices would occur, modelled on the individual R1 performances from the penultimate session. Turning again to our hypothesis that active inhibition of the exploratory drive is the key factor in gating sustained R1 expression, we suggest that mice, upon initial introduction to the radial maze, may not instantly engage this inhibition, thus allowing the exploratory drive to manifest more strongly at the beginning of the session. Moreover, the preferentially early per-session manifestation of exploratory activity in the R1 environment could be seen as refutational in nature; by

exploring S0 early, mice can be interpreted as testing the validity of R1, not only by confirming it (visiting S1) but also by actively ("active" because earlier than predicted) attempting to refute it (visiting S0). A consideration such as this, examined in the light of classical literature from the psychology of reasoning (Wason, 1960, 1966, 1968), prompts the intriguing evolutionary notion that mice may retain a stronger refutational reflex than adult humans.

Exploration better explained by global information gain than by foraging.

What Wason highlighted in his experiments is that, when reasoning on rules, attempting to refute the validity of a rule (e.g. "Test S0: if no reward, revert to R1 behavior") is a statistically more reliable means of maximizing information for the agent than attempting only to confirm it (e.g. "Test S1: if reward, stick with R1 behavior"). That we observed initial S0 choices occurring earlier than a S1 choice probability distribution could explain brings further evidence to the school of thought in behavioral psychology which places a prime on learning (in the sense of information gain), above reward foraging, as the primary driver of exploratory behavior, or what is commonly referred to as "curiosity" (Gaffan & Davies, 1981; Inglis et al., 2001; Kidd & Hayden, 2015). This is also interesting with respect to the classical dichotomy between goal-directed and habitual behavior. In our protocol, we observed clear goal-directed behavior (e.g. early initial S0 choice) even subsequent to S-R scores of 95 to 100%, scores which would normally be considered as signs that a habitual responding level was being reached. This proactive exploratory behavior would be difficult to explain if reward were the primary goal, but less so if we take the primary goal to be global information gain (of which reward location is but one element), especially in the context of large environments such as the radial maze. Looking to the literature, food restriction per se has previously been associated with increased exploratory behavior in rodents (Gelegen et al., 2006; Heinz et al., 2021). However, in these studies exploration is simply conflated with foraging behavior in a way that, as mentioned above, does not correspond with our observations. In fact, direct investigation of this precise question is largely lacking in the literature. In one recent study, where reward preference was established prior to a T-maze task, researchers found that the exploratory drive in non-food restricted mice led them to spatially alternate their arm choices more than they chose the reward containing arm of

the maze to which they had previously developed a preference (Habedank et al., 2021). In contrast, in another recent study, mice restricted to 85% of their baseline weight rapidly developed a strong side preference (~90%) in a free-choice T-maze task where they were rewarded regardless of the arm they chose (Park et al., 2021). A comparative reading of these studies seems, once again, to reveal a clear distinction between, on the one hand, exploration per se and, on the other hand, foraging behavior in the strict sense of food-directed behavior. Bearing in mind that the animals in our study were also food restricted to 85-90% baseline, when we further compare the rapidly established (3 sessions) side-preference observed in the T-maze apparatus (Park et al., 2021) to the more gradually attained sustained expression of R1 in the results from our present study in the larger radial maze apparatus (11+ sessions), this suggests a complex yet intuitive interaction between level of environmental uncertainty (low in the small surface area illuminated T-maze; high in the large surface area non-illuminated radial maze) and strength of information gain exploratory drive versus foraging drive. Once a hungry, foraging animal is satisfied it has minimized environmental uncertainty (i.e. in a highly simplistic environment), there will be much less motivation for it to continue exploring – which would require a certain cognitive budget via e.g. working memory – rather than repeatedly returning to a reliable food source by simply repeating a low cognitive cost proceduralized action. However, in the kind of natural environments in which rodents have evolved, environmental uncertainty would certainly be much higher than in a T-maze. Notably, with respect to our model, it also confirms that it would make little sense to speak of "indoctrination" in a learning environment where the innate exploratory drive is *not* maximized. With respect to all of these environmental considerations around exploratory drive persistence, future research could focus specifically on over-training mice towards R1 exploitation, studying how long it takes for S-R behavior to become truly procedural in mice in the larger, higher uncertainty, and in that sense therefore more naturalistic, radial maze apparatus. Another related prediction worth investigation is that post R1 criterion food reward devaluation, via satiety, would restore exploratory behavior, to the expense of R1 performance. Furthermore, while in this study we limited our indoctrination-like protocol reinforcers to presence and absence of reward only, future work could go further and study the impact of including a mild S0-aversion association (e.g. air-puff or quinine) alongside the S1-reward association. In the presence of such "punishment" of exploratory behavior, we should expect to observe a significant

acceleration in inhibition of the innate exploratory drive and, with this, accelerated and strengthened sustained expression of R1.

Direct pathway implication in "shift" and "stay" strategies.

The above considerations relating to exploration versus exploitation in the context of the radial maze contributed to a broader line of reflection regarding the nature of what the literature refers to as "win-shift" versus "win-stay" behaviors (Gaffan & Davies, 1981; Packard et al., 1989; Sage & Knowlton, 2000). In reviewing literature from both maze and operant protocols, it struck us that the cognitive implications, from the animal's perspective, of engaging in either win-shift or win-stay behavior should be strongly dependent upon whatever the innate, spontaneous behavior of the animal would be in a given environment (i.e. maze or operant). Exploring this question was important to us because of recent studies investigating the specific direct and indirect pathway activations correlated with win-shift versus win-stay behavior, work which, crucially, was conducted only in operant conditions (Kwak & Jung, 2019; Nonomura et al., 2018; Vicente et al., 2016). This work showed that the D_1-expressing direct pathway codes for "reward" signals and the consequent decision to "stick" to the current operandum, whereas the D_2-expressing indirect pathway codes for "no-reward" signals and the consequent decision to "switch" to the other operandum. However, convergent evidence from both published studies (Reed, 2016; Vannoni et al., 2014) and unpublished observations discussed in correspondence with several researchers (Allen Neuringer, Chris Rodgers, Jonny Saunders) strongly suggests that in operant conditions (levers, lickers, or nose-pokes) "stay" is the statistically more likely spontaneous baseline action in rodents. This is in marked contrast to maze conditions where it is well established that "shift" is the spontaneously dominant strategy. Based on this contextual baseline discrepancy, we proceeded to reframe the findings from the operant literature and instead hypothesize that direct pathway activity may be correlated to whichever action corresponds closest to an animal's statistically most likely spontaneous action in a given environment, i.e. "stay" in operant conditions but "shift" in maze conditions. Consequently, if our hypothesis were accurate, then potentiating the intrinsic activity of the direct pathway in the R1 maze environment should in turn potentiate exploratory

behavior, making it more difficult to inhibit, and thus more difficult for R1 exploitatory behavior to be expressed.

CB_1-mediated inhibitory control in direct pathway and its behavioral implications.

To test the direct-pathway potentiation hypothesis, we turned to the D_1-CB_1-KO transgenic mouse line, in which the CB_1 receptor is conditionally deleted from all D_1-positive neurons of the forebrain (Monory et al., 2007), including the medium spiny neurons composing the direct pathway in the dorsal striatum, but not including the D_2-positive neurons composing the indirect pathway.

That this D_1-CB_1-KO population displayed lower R1 expression than their wildtype littermates in both iterations of the experiment, can be taken as evidence towards our cognitive interpretation of how the direct and indirect pathways flexibly interact as acquired exploitatory behaviors compete for expression with the innate exploratory drive. Notably, inhibitory modulatory control of the direct pathway provided by CB_1 seems to play a significant role. Regarding our hypothesis that R1 expression but not acquisition would be affected by manipulation of the direct pathway, when we divided the D_1-CB_1-KO population according to whether or not R1 criterion performance had been reached within 13 sessions, we nevertheless observed simultaneous emergence of comparably greater run times on S0 vs S1 in both below criterion and above criterion animals. Below criterion animals did, however, perform significantly less rectifying choice revision behavior compared to their above criterion D_1-CB_1-KO littermates. This run time versus choice revision phenotype discrepancy in below criterion animals supports our general interpretation that successful exploitation of R1, in both C57Bl6/J and D_1-CB_1-KO animals alike, is achieved via an active process of inhibiting the persistent innate exploratory drive, implying that the more likely *initial* strategy choice on a given trial, even long after the S1-reward association has been acquired, remains spontaneous exploration rather than R1 exploitation. This initial strategy selection would then have to be actively over-ruled, i.e. revised, putatively via a combination of indirect pathway-mediated (Nonomura et al., 2018) and retrograde CB_1-mediated signals. If direct pathway signals are potentiated by CB_1 deletion, as has been previously shown (Soria-Gomez et al., 2021), then correction of initial movement-initiating activity would be rendered more difficult, constituting a strong neurobiological candidate for

the phenotype we observed. This exploration priming interpretation of our trial-by-trial observations is also compatible with the observation discussed above that initial session-by-session S0 exploration occurs preferentially in the earliest trials, even in the final R1 training session. We also gathered further complementary evidence for this interpretation when we conducted R1 training with Dlx-CB_1-KO animals who lack CB_1 on all GABAergic neurons of the forebrain, i.e. including both the direct and indirect pathways. If we can posit that this "twin" potentiation re-establishes a new equilibrium in the competition between direct and indirect pathway mediated strategies, then this provides a strong candidate explanation for why we did not observe a decrease in R1 expression in Dlx-CB_1-KO mice compared to their wildtype littermates. Future investigation into our direct pathway potentiation hypothesis could make use of cued optogenetic approaches which would allow for increased direct pathway activity on specific trials or arm pairs but not others, thereby enabling within-subject performance comparison.

To verify that restoring modulatory inhibitory control to the direct pathway, via CB_1, would be sufficient to restore the inhibitory striatal flexibility necessary for the innate exploratory strategy to be overcome by the acquired exploitatory one, we virally re-expressed CB_1 receptors locally in D_1-positive neurons of the dorsal striatum of D_1-CB_1-KO mice. Here, we observed that D_1-CB_1-$KO_{Str+/+}$ animals expressed the R1 rule significantly stronger and more robustly (i.e. with less variance) than their D_1-CB_1-$KO_{Str-/-}$ littermates who had been injected with an empty viral vector. As mentioned in the results, R1 exploitation in the D_1-CB_1-$KO_{Str+/+}$ group even tended to be stronger and more robust than wildtype performances. This latter result we tentatively attribute to our viral reexpression giving rise to a relative *over*-expression of CB_1 (see Materials & Methods), and thus increased direct pathway inhibitory control, compared to wildtype animals. In either case, the result reveals that adding inhibitory control to the direct pathway facilitates expression of an exploitatory S-R rule, enabling the flexibility needed to inhibit and switch away from innate responses. Relative to previous work showing that CB_1 expressed on D_1-positive neurons in the hippocampus play a role in novel-object recognition memory (Oliveira da Cruz et al., 2020), this result with local CB_1 reexpression in the striatum only also demonstrates that the hippocampal D_1-CB_1 population is not necessary for successful and timely expression of an exploitatory tactile-discrimination rule. This also corroborates what we observed when we compared

(in young C57Bl6/J mice) results from an experimental set-up with a fixed tactile spatial map to one in which the map was fluid, i.e. modified every session: no difference in R1 performance between the fixed and fluid groups, indicating that a hippocampal spatial strategy was not being used even when putatively available.

Aged mice display decreased inhibition of exploration strategy.

Even more than the D_1-CB_1-KO population, aged animals were also highly impaired in their ability to overcome the innate exploratory drive in order to exploit the R1 strategy, but again despite no observable impairment in their acquisition of the S1-reward association. Indeed, in the present study certain aged mice, trained for longer than the others yet still never reaching R1 criterion, nevertheless displayed the same characteristic S1-reward acquisition phenotypes of robust differential S0 vs S1 run time plus significantly more rectifying than error-inducing choice revision. If exploratory responses do develop into some kind of innate "chunk" strategy in a given environment, as hypothesized above, then, as cognitive flexibility in the form of active inhibition decreases with age, increasingly rigid cognitive selection of such a "chunk" is what we should expect to observe, despite unimpaired acquisition of the reward-association basis for a competing strategy. Since this well-characterized age-related decline in inhibitory cognitive flexibility (Coxon et al., 2012) correlates with, among other things, a decrease in levels of both CB_1 and D_1 (Bilkei-Gorzo, 2012; Wang et al., 1998), it would be interesting in future work to over-express CB_1 in the direct pathway of aged mice, on the hypothesis that this would improve their capacity to switch from innate exploratory to acquired exploitatory strategies. Indeed, recent work has shown that chronic treatment of aged mice with THC restores subsequent flexible cognitive function (Bilkei-Gorzo et al., 2017), demonstrating that potentiation of the endocannabinoid system has therapeutic value in the reversal of the cognitive impacts of ageing.

Taken together, the results from our experiment with aged mice further illustrate the persistence of the innate, evolutionarily preserved drive towards exploration in mice, revealing that, with age, deterioration of inhibitory cognitive flexibility seems to translate into reinforced rigidity in the selection of primitive, unrewarded exploratory strategies, despite robust acquisition of competing, rewarded strategies. Bringing all this back to the objective of our study, it appears that as inhibitory cognitive flexibility decreases with age or genetic manipulation, resistance to indoctrination increases, though the mechanism underpinning this resistance may more accurately be described as a reduced capacity to flexibly inhibit competing innate behaviors. As a final thought we do not have the scope to develop here, the phenotypes we observed across our experimental groups are highly reminiscent of those seen in tasks which also rely on a specific type of inhibitory cognitive flexibility referred to as adaptive forgetting (Anderson & Floresco, 2021; Bekinschtein et al., 2018; Schmitz et al., 2017; Hulbert et al., 2016). While the literature on adaptive forgetting has thus far been focused on hippocampal memory functions, generalizing the striatal results we have observed in our task into this same explanatory paradigm may prove very fruitful to future research avenues. Within this paradigm, indoctrination would be a hijacking of adaptive forgetting as a means of suppressing "unwanted thoughts" (Schmitz et al., 2017) with

***Differential run time*:** Run time has previously been analyzed by our laboratory as a parameter in a radial maze "Go/No-Go" protocol (Marighetto et al., 1999). In such experiments, animals sometimes "go" even when they have learnt there will be no reward at the location of the presented option, i.e. on arms where they should "No-Go". However, when trained animals "go" on a "No-Go" trial, their run time advancing towards the reward zone of the arm is higher than on "Go" trials. In the present study, we present run time as a proxy measure of post-choice confidence. Several justifications for this interpretation can be found in our observations. Firstly, comparing run times between the two C57Bl/6J populations during R1 training, these were lower in always-rewarded control animals on both S0 and S1 arms than they were in trained animals even on S1 arms. This indicates that animals in the always-rewarded control group correspondingly displayed an overall, surface independent, higher level of per trial post-choice confidence in finding a reward. Secondly, in classical declarative memory experiments (see Stevens et al., 2022b), run time was significantly lower on correct choice trials compared to incorrect trials, especially on trials of lower complexity, indicating that reward location confidence shapes run time differences more than tactile differences per se. This differential run time phenotype may also have an affective dimension, putatively related to amygdalar function (McDonald et al., 2004; McDonald and Hong, 2004).

Reliable cognitive measures for post-choice confidence in rodent models, especially mice, are relatively lacking in the literature (Carandini and Churchland, 2013; Hanks and Summerfield, 2017; Kepecs et al., 2008), despite this being a fundamental component of decision making processes. In the radial maze, this component is gained as a fact of the apparatus itself, enabling analysis of cognitive processes in mice while they are physically realizing their choice, in a manner not possible where execution is quasi-instantaneous (e.g. lever press). This opens up exciting possibilities for future *in vivo* investigation into its neurobiological bases.

the sole precision that it would be the educator who has decided which thoughts are unwanted.

Choice revision: Another important decision-making relevant cognitive behavior afforded by the radial maze apparatus is physically manifest choice revision, whereby an animal crosses, partially or fully, the threshold of one arm of the presented pair in a trial but then stops, turns around, and returns to the central platform to revise its choice, either by entering the other arm of the current pair or (less frequently) by re-entering the same arm, a process which can even be repeated several times within a given trial. Like run time as a reflection of post-choice confidence, choice revision behavior presents an exciting prospect for *in vivo* investigation. With only minor tweaking, it would be relatively easy to identify, from live video tracking of a session, the instant at which an animal physically turns back on its most recent choice and then correlate brain activity recorded *in vivo* to this instant in a classical peri-event manner. The experimental and control groups combined (n = 78) performed a total of 2,584 manifest choice revisions, or identifiable 'KOOK' events, during R1 training with the former group manifesting around twice as much choice revision as control animals, making of such events robust and fertile ground for future research.

Relative to the existing literature in decision-making, choice revision in the present study is similar to what is referred to as "vicarious trial-and-error" (VTE), first identified and theorized in now classic papers in behavioral psychology (Muenzinger and Gentry, 1931; Tolman, 1948, 1939). An excellent review of classical and recent work into the psychology and neurobiology of VTE can be found in (Redish, 2016). However, there are important differences between classical VTE and what we call choice revision here. On a technical level, VTE, most often studied in rats, is ethologically more fine-grained since it quantifies not only full-body movements but also head movements accompanying an animal looking back and forth. Our mice also made such movements, but at the scale of the radial maze apparatus the sensitivity of our motion tracking equipment allowed us to identify and quantify only large or full-body movements. For this reason, we suggest that the behavior we report here as choice revision is a representative fraction only of the full range of cognitive choice revision occurring both mentally and in a range of discrete to large physical movements. Full-body choice revision does carry the advantage of implying a temporal sequence in which action execution, action revision, action arrest, alternative action execution, etc., can be isolated. Pursuant to this, it is interesting to note on a cognitive level that, contrary to the literature on VTE, we did not observe a decrease in choice revision as tactile S-R behavior shifted to being putatively more procedural. As we have already discussed, in the context of the radial maze, the exploratory drive in mice is tenaciously persistent, both in the results we report here and in previous studies from our laboratory (Marighetto et al., 1999). As a result, we observed high levels of rectifying choice revision (correcting an initial exploratory action to an exploitatory one) even after 200+ trials.

Concluding remarks.

It is clearly not plausible to imagine that mammals have adaptively evolved mechanisms for the purpose of resisting indoctrination. Rather, the experiments we have run with the present model strongly suggest that the conditions for indoctrination are primarily a function (or artefact) of our species' unique environment, and therefore not primarily a function of our species' unique neural make-up. Human indoctrination, an eminently social phenomenon, may therefore constitute an instance of social cooptation of evolutionarily much older mechanisms of neurocognitive flexibility: an artefact of human socio-psychological environments in which unchecked exploratory behavior is often perceived as being disadvantageous to a given population. This stands counter to recent literature which has rejected the possibility that cognitive mechanisms analogous to indoctrination or confirmation bias could exist in non-human mammals (Mercier & Sperber, 2017).

In the present paper, we have focused on those mechanisms involved *during* indoctrination, while in a follow-up paper to this one (Stevens et al., 2022b) we focus on the cognitive consequences of entering a new environment which demands that the content of this indoctrination be revised. We believe we have demonstrated the validity of this initial indoctrination phase of the model, especially with respect to the clear and lasting disambiguation it elicits between, on the one hand, acquisition of the basic S1-reward association and, on the other hand, sustained exploitatory expression of it. Beyond the clear pedagogical, political, and social interest of gaining a deeper cognitive and neurobiological understanding of how the mammalian mind-brain bends under such situations of indoctrination, the introduction of this behavioral model also opens several avenues for future pre-clinical research in domains having known cognitive interactions, either as cause or effect, with indoctrination such as addiction (Norton, 1994), psychological disorders including delusions and schizophrenia (Wareham, 2019), trauma recovery (Curtis & Curtis, 1993), and others. Finally, the validation of this model alters the research environment itself with respect to our current thinking around the neural and environmental evolution of the higher reasoning faculties of humans (for discussion of this angle specifically, see Stevens 2022), adding a new layer to our

considerations of the broader neural, cognitive, social, and ultimately ecological implications of how we learn and how we teach.

Bibliography

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological Seduction and Self-Radicalization. *Journal of the American Philosophical Association*, *4*(3), 298–322. https://doi.org/10.1017/apa.2018.27

Anderson, M. C., & Floresco, S. B. (2021). Prefrontal-hippocampal interactions supporting the extinction of emotional memories: The retrieval stopping model. *Neuropsychopharmacology*, 1–16. https://doi.org/10.1038/s41386-021-01131-1

Bedre, R. (2021). *reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit.* Zenodo. https://doi.org/10.5281/zenodo.4422035

Bekinschtein, P., Weisstaub, N. V., Gallo, F., Renner, M., & Anderson, M. C. (2018). A retrieval-specific mechanism of adaptive forgetting in the mammalian brain. *Nature Communications*, *9*(1), 4660. https://doi.org/10.1038/s41467-018-07128-7

Bilkei-Gorzo, A. (2012). The endocannabinoid system in normal and pathological brain ageing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1607), 3326–3341. https://doi.org/10.1098/rstb.2011.0388

Bilkei-Gorzo, A., Albayram, O., Draffehn, A., Michel, K., Piyanova, A., Oppenheimer, H., Dvir-Ginzberg, M., Rácz, I., Ulas, T., Imbeault, S., Bab, I., Schultze, J. L., & Zimmer, A. (2017). A chronic low dose of Δ9-tetrahydrocannabinol (THC) restores cognitive function in old mice. *Nature Medicine*, *23*(6), 782–787. https://doi.org/10.1038/nm.4311

Callan, E., & Arena, D. (2009, October 30). *Indoctrination.* The Oxford Handbook of Philosophy of Education. https://doi.org/10.1093/oxfordhb/9780195312881.003.0007

Carland, M. A., Thura, D., & Cisek, P. (2019). The Urge to Decide and Act: Implications for Brain Function and Dysfunction. *The Neuroscientist*, *25*(5), 491–511. https://doi.org/10.1177/1073858419841553

Cisek, P., & Hayden, B. Y. (2022). Neuroscience needs evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200518. https://doi.org/10.1098/rstb.2020.0518

Cohen, J., Westlake, K., & Szelest, I. (2004). Effects of runway shift and stay rules on rats' serial pattern learning in the T-maze. *Animal Learning & Behavior*, *32*(4), 500–511. https://doi.org/10.3758/BF03196045

Colombo, P., Davis, H., & Volpe, B. (1990). Allocentric Spatial and Tactile Memory Impairments in Rats With Dorsal Caudate Lesions Are Affected by Preoperative Behavioral Training. *Behavioral Neuroscience*, *103*, 1242–1250. https://doi.org/10.1037//0735-7044.103.6.1242

Conant, J. B. (1948). Education in a Divided World: The Function of the Public School in Our Unique Society. In *Education in a Divided World*. Harvard University Press. https://doi.org/10.4159/harvard.9780674283756

Coxon, J. P., Impe, A. V., Wenderoth, N., & Swinnen, S. P. (2012). Aging and Inhibitory Control of Action: Cortico-Subthalamic Connection Strength Predicts Stopping Performance. *Journal of Neuroscience*, *32*(24), 8401–8412. https://doi.org/10.1523/JNEUROSCI.6360-11.2012

Crozier, M., Huntington, S. P., & Watanuki, J. (1975). *The crisis of democracy: Report on the governability of democracies to the Trilateral Commission*. New York University Press.

Curtis, J. M., & Curtis, M. J. (1993). Factors Related to Susceptibility and Recruitment by Cults. *Psychological Reports*, *73*(2), 451–460. https://doi.org/10.2466/pr0.1993.73.2.451

Dehaene, S., & Cohen, L. (2007). Cultural Recycling of Cortical Maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dember, W. N., & Richman, C. L. (1989). *Spontaneous Alternation Behavior*. Springer New York. https://doi.org/10.1007/978-1-4613-8879-1

Dewey, J. (1916). *Democracy and Education*. Southern Illinois University Press ; Feffer & Simons.

Dhawale, A. K., Wolff, S. B. E., Ko, R., & Ölveczky, B. P. (2021). The basal ganglia control the detailed kinematics of learned motor skills. *Nature Neuroscience*, *24*(9), 1256–1269. https://doi.org/10.1038/s41593-021-00889-3

Dudman, J. T., & Krakauer, J. W. (2016). The basal ganglia: From motor commands to the control of vigor. *Current Opinion in Neurobiology*, *37*, 158–166. https://doi.org/10.1016/j.conb.2016.02.005

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Gaffan, E. A., & Davies, J. (1981). The role of exploration in win-shift and win-stay performance on a radial maze. *Learning and Motivation*, *12*(3), 282–299. https://doi.org/10.1016/0023-9690(81)90010-2

Gelegen, C., Collier, D. A., Campbell, I. C., Oppelaar, H., & Kas, M. J. H. (2006). Behavioral, physiological, and molecular differences in response to dietary restriction in three inbred mouse strains. *American Journal of Physiology-Endocrinology and Metabolism*, *291*(3), E574–E581. https://doi.org/10.1152/ajpendo.00068.2006

Gould, S. J., Lewontin, R. C., Maynard Smith, J., & Holliday, R. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *205*(1161), 581–598. https://doi.org/10.1098/rspb.1979.0086

Gould, S. J., & Vrba, E. S. (1982). Exaptation—A Missing Term in the Science of Form. *Paleobiology*, *8*(1), 4–15. https://doi.org/10.1017/S0094837300004310

Graybiel, A. M. (1998). The Basal Ganglia and Chunking of Action Repertoires. *Neurobiology of Learning and Memory*, *70*(1), 119–136. https://doi.org/10.1006/nlme.1998.3843

Habedank, A., Kahnau, P., & Lewejohann, L. (2021). Alternate without alternative: Neither preference nor learning explains behaviour of C57BL/6J mice in the T-maze. *Behaviour*, *158*(7), 625–662. https://doi.org/10.1163/1568539X-bja10085

Heinz, D. E., Schöttle, V. A., Nemcova, P., Binder, F. P., Ebert, T., Domschke, K., & Wotjak, C. T. (2021). Exploratory drive, fear, and anxiety are dissociable and independent components in foraging mice. *Translational Psychiatry*, *11*(1), 1–12. https://doi.org/10.1038/s41398-021-01458-9

Hitoshi, N., Ken-ichi, Y., & Jun-ichi, M. (1991). Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene*, *108*(2), 193–199. https://doi.org/10.1016/0378-1119(91)90434-D

Houdé, O. (2019). *3-System Theory of the Cognitive Brain: A Post-Piagetian Approach to Cognitive Development*. Routledge. https://doi.org/10.4324/9781315115535

Hulbert, J. C., Henson, R. N., & Anderson, M. C. (2016). Inducing amnesia through systemic suppression. *Nature Communications*, *7*(1), 11003. https://doi.org/10.1038/ncomms11003

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Inglis, I. R., Langton, S., Forkman, B., & Lazarus, J. (2001). An information primacy model of exploratory and foraging behaviour. *Animal Behaviour*, *62*(3), 543–557. https://doi.org/10.1006/anbe.2001.1780

Jin, X., Tecuapetla, F., & Costa, R. M. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nature Neuroscience*, *17*(3), 423–430. https://doi.org/10.1038/nn.3632

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*(1), 39589. https://doi.org/10.1038/srep39589

Kidd, C., & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Kuhn, T. S. (1977). *The Essential Tension Selected Studies in Scientific Tradition and Change*. University of Chicago Press.

Kuhn, T. S. (1961). The Function of Dogma in Scientific Research-pp. 347-69 in AC Crombie (ed.). Scientific Change. *Symposium on the History of Science University of Oxford*, 9–15.

Kwak, S., & Jung, M. W. (2019). Distinct roles of striatal direct and indirect pathways in value-based decision making. *ELife*, *8*, e46050. https://doi.org/10.7554/eLife.46050

Lalonde, R. (2002). The neurobiological basis of spontaneous alternation. *Neuroscience & Biobehavioral Reviews*, *26*(1), 91–104. https://doi.org/10.1016/S0149-7634(01)00041-0

Levins, R., & Lewontin, R. (1985). *The Dialectical Biologist*. Harvard University Press.

Lewontin, R., & Levins, R. (2000). Schmalhausen's law. *Capitalism Nature Socialism*, *11*(4), 103–108. https://doi.org/10.1080/10455750009358943

Lorenz, K. Z. (1958). The evolution of behavior. *Scientific American*, *199*(6), 67-74 passim. https://doi.org/10.1038/scientificamerican1258-67

Macmillan, C. J. B. (1983). On Certainty and indoctrination. *Synthese*, *56*(3), 363–372. https://doi.org/10.1007/BF00485472

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Marighetto, A., Etchamendy, N., Touzani, K., Torrea, C. C., Yee, B. K., Rawlins, J. N. P., Jaffard, R., & Marighetto, Aline. (1999). Knowing which and knowing what: A potential mouse model for age-related human declarative memory decline: Mouse model for memory decline. *European Journal of Neuroscience*, *11*(9), 3312–3322. https://doi.org/10.1046/j.1460-9568.1999.00741.x

Marighetto, A., Touzani, K., Etchamendy, N., Torrea, C. C., De Nanteuil, G., Guez, D., Jaffard, R., & Morain, P. (2000). Further Evidence for a Dissociation Between Different Forms of Mnemonic Expressions in a Mouse Model of Age-related Cognitive Decline: Effects of Tacrine and S 17092, a Novel Prolyl Endopeptidase Inhibitor. *Learning & Memory*, *7*(3), 159–169. https://doi.org/10.1101/lm.7.3.159

McDonald, R. J., & Hong, N. S. (2004). A dissociation of dorso-lateral striatum and amygdala function on the same stimulus–response habit task. *Neuroscience*, *124*(3), 507–513. https://doi.org/10.1016/j.neuroscience.2003.11.041

McDonald, R. J., & White, N. M. (1993). *A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. - PsycNET.* https://content.apa.org/doiLanding?doi=10.1037%2F0735-7044.107.1.3

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason.* Harvard University Press.

Monory, K., Blaudzun, H., Massa, F., Kaiser, N., Lemberger, T., Schütz, G., Wotjak, C. T., Lutz, B., & Marsicano, G. (2007). Genetic Dissection of Behavioural and Autonomic Effects of Δ9-Tetrahydrocannabinol in Mice. *PLoS Biology*, *5*(10), e269. https://doi.org/10.1371/journal.pbio.0050269

Munro, M. (2015). The hijacked brain. *Nature*, *522*(7557), S46–S47. https://doi.org/10.1038/522S46a

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nonomura, S. (2018). *Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways.* 19.

Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., Watanabe, M., Yamanaka, K., Enomoto, K., Chiken, S., Sano, H., Soma, S.,

Yoshida, J., Samejima, K., Ogawa, M., Kobayashi, K., Nambu, A., Isomura, Y., & Kimura, M. (2018). Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. *Neuron*, *99*(6), 1302-1314.e5. https://doi.org/10.1016/j.neuron.2018.08.002

Norton, J. H. (1994). Addiction and family issues. *Alcohol*, *11*(6), 457–460. https://doi.org/10.1016/0741-8329(94)90066-3

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Oliveira da Cruz, J. F., Busquets-Garcia, A., Zhao, Z., Varilh, M., Lavanco, G., Bellocchio, L., Robin, L., Cannich, A., Julio-Kalajzić, F., Lesté-Lasserre, T., Maître, M., Drago, F., Marsicano, G., & Soria-Gómez, E. (2020). Specific Hippocampal Interneurons Shape Consolidation of Recognition Memory. *Cell Reports*, *32*(7), 108046. https://doi.org/10.1016/j.celrep.2020.108046

Olton, D. S., & Samuelson, R. J. (1976). Remembrance of places passed: Spatial memory in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*(2), 97–116. https://doi.org/10.1037/0097-7403.2.2.97

Packard, M., Hirsh, R., & White, N. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *The Journal of Neuroscience*, *9*(5), 1465–1472. https://doi.org/10.1523/JNEUROSCI.09-05-01465.1989

Park, A. J., Harris, A. Z., Martyniuk, K. M., Chang, C.-Y., Abbas, A. I., Lowes, D. C., Kellendonk, C., Gogos, J. A., & Gordon, J. A. (2021). Reset of hippocampal–prefrontal circuitry facilitates learning. *Nature*, *591*(7851), 615–619. https://doi.org/10.1038/s41586-021-03272-1

Popper, K. R. (1935). *The Logic of Scientific Discovery*. Routledge. http://nukweb.nuk.uni-lj.si/login?url=http://search.ebscohost.com/login.aspx?authtype=ip&direct=true&db=nlebk&AN=143035&site=eds-live&scope=site&lang=sl

Reback, J., McKinney, W., jbrockmendel, Bossche, J. V. den, Augspurger, T., Cloud, P., gfyoung, Sinhrks, Klein, A., Roeschke, M., Hawkins, S., Tratner, J., She, C., Ayd, W., Petersen, T., Garcia, M., Schendel, J., Hayden, A., MomIsBestFriend, … Mehyar, M. (2020). *pandas-dev/pandas: Pandas 1.0.3*. Zenodo. https://doi.org/10.5281/zenodo.3715232

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17*(3), 147–159. https://doi.org/10.1038/nrn.2015.30

Reed, P. (2016). Win-stay and win-shift lever-press strategies in an appetitively reinforced task for rats. *Learning & Behavior*, *44*(4), 340–346. https://doi.org/10.3758/s13420-016-0225-2

Reid, A. K., Dixon, R., & Gray, S. (2008). Variation and selection in response structures. In *Reflections on adaptive behavior: Essays in honor of J. E. R. Staddon* (pp. 51–85). MIT Press.

Reynolds, C. M., & Canna, S. (2012). *Topics for Operational Considerations: Insights from Neurobiology & Neuropsychology on Influence and Extremism—An Operational Perspective*. 86.

Richman, C. L., Dember, W. N., & Kim, P. (1986). Spontaneous alternation behavior in animals: A review. *Current Psychological Research & Reviews*, *5*(4), 358–391. https://doi.org/10.1007/BF02686603

Sage, J., & Knowlton, B. (2000). Effects of US devaluation on win-stay and win-shift radial maze performance in rats. Behav. Neurosci. 114, 295-306. *Behavioral Neuroscience*, *114*, 295–306. https://doi.org/10.1037/0735-7044.114.2.295

Schmalhausen, I. I. (1949). *Factors of evolution: The theory of stabilizing selection* (pp. xiv, 327). Blakiston.

Schmitz, T. W., Correia, M. M., Ferreira, C. S., Prescot, A. P., & Anderson, M. C. (2017). Hippocampal GABA enables inhibitory control over unwanted thoughts. *Nature Communications*, *8*(1), 1311. https://doi.org/10.1038/s41467-017-00956-z

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23–32.

Soria-Gomez, E., Zottola, A. C. P., Mariani, Y., Desprez, T., Barresi, M., Río, I. B., Muguruza, C., Bon-Jego, M. L., Julio-Kalajzić, F., Flynn, R., Terral, G., Fernández-Moncada, I., Robin, L. M., Cruz, J. F. O. da, Corinti, S., Amer, Y. O., Goncalves, J., Varilh, M., Cannich, A., … Bellocchio, L. (2021). Subcellular specificity of cannabinoid effects in striatonigral circuits. *Neuron*, *109*(9), 1513-1526.e11. https://doi.org/10.1016/j.neuron.2021.03.007

Staddon, J. E., & Simmelhag, V. L. (1971). The "supersitition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psychological Review*, *78*(1), 3–43. https://doi.org/10.1037/h0030305

Stevens, C., Lacroix, C., Bouchet, M., Mariani, Y., Roudier, F., Marsicano, G., & Marighetto, A. (2022b). Investigating hallmarks of 'myside' confirmation-bias like behaviors in a novel mouse model of rule revision. (In preparation)

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). A Bradford Book.

Taylor, R. M. (2017). Indoctrination and Social Context: A System-based Approach to Identifying the Threat of Indoctrination and the Responsibilities of Educators. *Journal of Philosophy of Education*, *51*(1), 38–58. https://doi.org/10.1111/1467-9752.12180

Tolman, E. C. (1939). *Prediction of vicarious trial and error by means of the schematic sowbug.* - *PsycNET.* https://psycnet.apa.org/doiLanding?doi=10.1037%2Fh0057054

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual.* CreateSpace.

Vannoni, E., Voikar, V., Colacicco, G., Sánchez, M. A., Lipp, H.-P., & Wolfer, D. P. (2014). Spontaneous behavior in the social homecage discriminates strains, lesions and mutations in mice. *Journal of Neuroscience Methods*, *234*, 26–37. https://doi.org/10.1016/j.jneumeth.2014.04.026

Vicente, A. M., Galvão-Ferreira, P., Tecuapetla, F., & Costa, R. M. (2016). Direct and indirect dorsolateral striatum pathways reinforce different action strategies. *Current Biology*, *26*(7), R267–R269. https://doi.org/10.1016/j.cub.2016.02.036

Wang, Y., Chan, G. L. Y., Holden, J. E., Dobko, T., Mak, E., Schulzer, M., Huser, J. M., Snow, B. J., Ruth, T. J., Calne, D. B., & Stoessl, A. J. (1998). Age-dependent decline of dopamine D1 receptors in human brain: A PET study. *Synapse*, *30*(1), 56–61. https://doi.org/10.1002/(SICI)1098-2396(199809)30:1<56::AID-SYN7>3.0.CO;2-J

Wareham, R. J. (2019). Indoctrination, delusion and the possibility of epistemic innocence. *Theory and Research in Education*, *17*(1), 40–61. https://doi.org/10.1177/1477878518812033

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

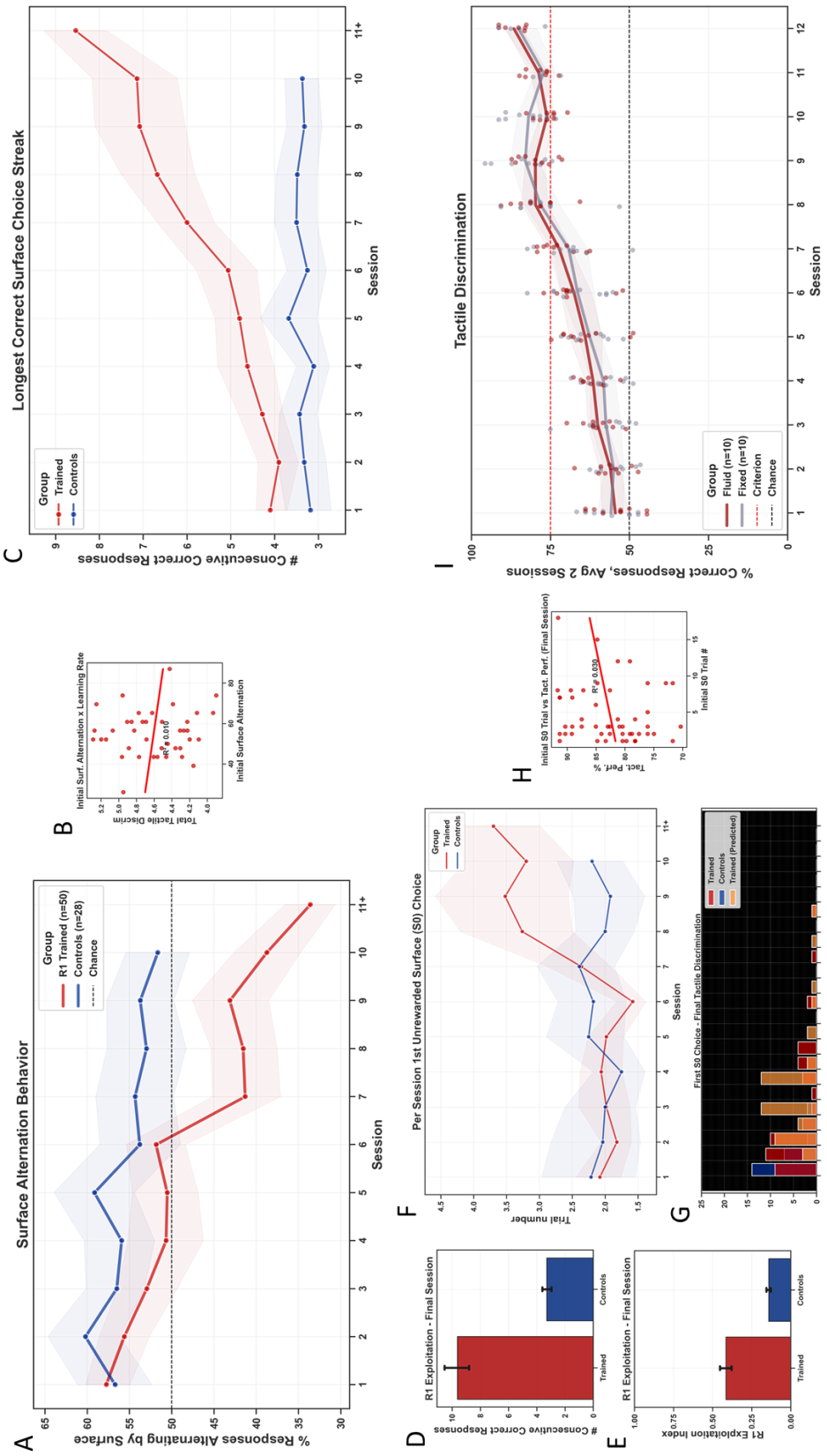
Wason, P. C. (1960). On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Wason, P. C. (1966). *New Horizons in Psychology*. Penguin Books.

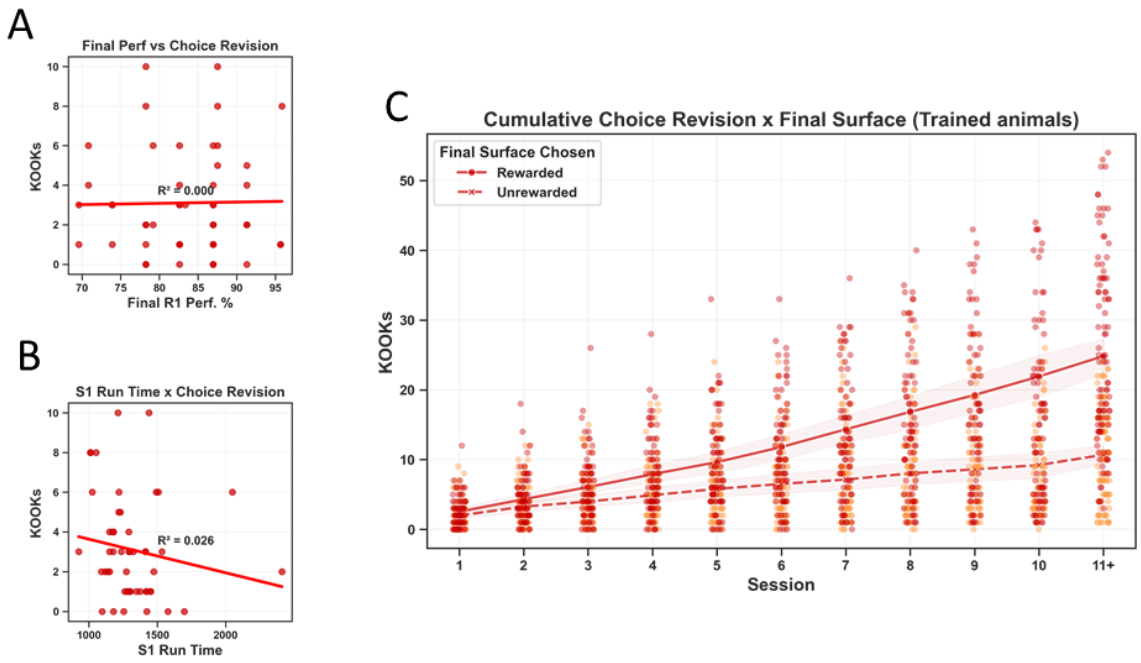
Wason, P. C. (1968). Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wexler, B. E. (2011). Neuroplasticity: Biological Evolution's Contribution to Cultural Evolution. In S. Han & E. Pöppel (Eds.), *Culture and Neural Frames of Cognition and Communication* (pp. 1–17). Springer. https://doi.org/10.1007/978-3-642-15423-2_1

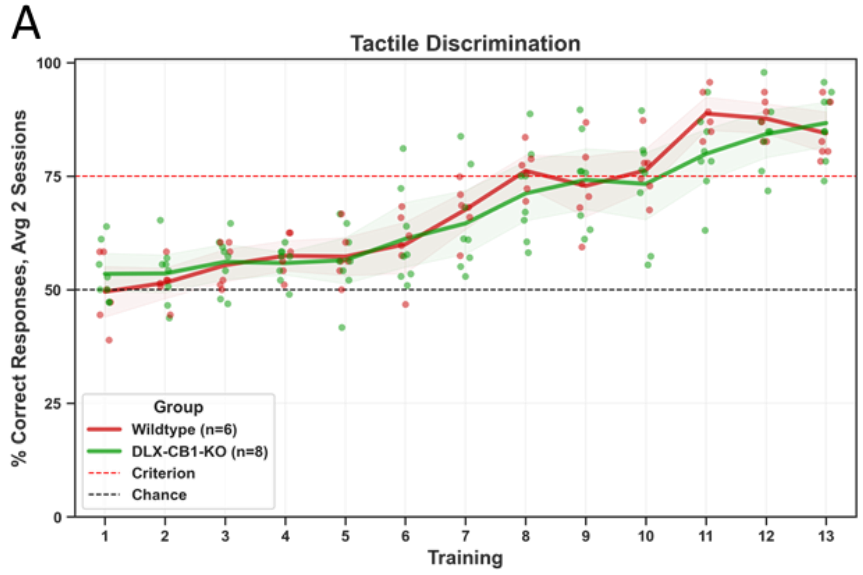
Supplementary Figures:



Supplementary figure S. 1 - Exploratory resistance to R1 expression.
Experimental animals are represented in red, controls in blue. All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. (**A**) Session-by-session surface alternation behavior reveals that control animals, rewarded on every trial regardless of surface chosen, tended to maintain above chance level alternation by surface behavior throughout R1 training. (**B**) Linear regression analysis revealed no correlation between individual strength of initial surface alternation in R1 trained animals and subsequent R1 expression performance, $R^2 = 0.01$. (**C**), (**D**) + (**E**) Looking at average longest run of consecutive correct S1 choices, while R1 trained animals quickly rise highly significantly above control animals, even in the final session these runs represent less than half the total number of trials in the session. (**F**) Despite clear R1 acquisition and increasing expression, in terms of on which trial per session animals first choose the exploratory S0 option, the R1 trained population chooses S0 significantly later than controls only from session 8 onwards, and with very large within population variance. (**G**) Based on individual R1 expression performance in the penultimate session, we modeled when R1 trained animals would be predicted to choose S0 in the final session (orange bars). This revealed that R1 trained animals (red bars) were in fact actively first making exploratory S0 choices earlier than could have been predicted by R1 performance only. Controls (blue bars), as expected, first chose S0 according to a random compatible distribution. (**H**) Using linear regression analysis, we nevertheless found no correlation between timing of initial exploratory S0 choice and subsequent R1 performance, $R^2 = 0.03$. (**I**) Testing to see if the possibility of a hippocampal dimension would impact R1 expression, we compared performance in one cohort where the tactile configuration changed in every session ('Fluid' group, n=10) to performance in a cohort for whom the tactile configuration remained the same in every session ('Fixed', n=10). No difference in R1 expression was observed, indicating that either a tactile map was not formed, or if it was, it neither aided nor impaired R1 expression. Both cohorts were thus subsequently pooled into the overall R1 trained population.

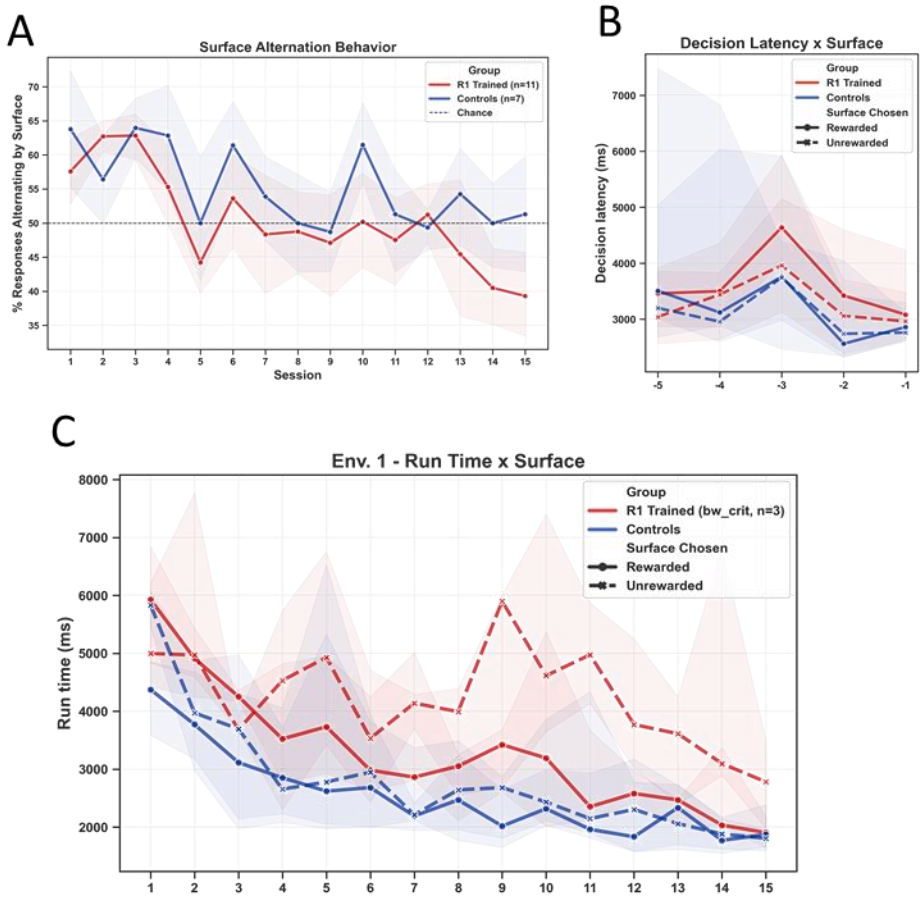


Supplementary figure S. 2 - Analysis of choice revision behavior.
(**A**) + (**B**) Since the existing literature on vicarious trial & error (VTE) predicts a decrease of deliberative behavior as a function of response proceduralization, we compared levels of choice revision in the final R1 session, where performances were highest and therefore closest to being proceduralized. However, linear regression analysis revealed no correlation between either final R1 performance and choice revision ($R^2 = 0$) or S1 run time and choice revision ($R^2 = 0.026$). (**C**) Cumulative analysis of choice revision behavior in R1 trained animals revealed very large inter-individual variance in this behavior, which was not, however, as the linear regressions in (A) and (B) show, related to either R1 performance or to run time.



Supplementary figure S. 3 - Deletion of CB1 receptors from all GABAergic neurons of forebrain does not negatively impact R1 expression.

Error bands represent 95% confidence intervals. (A) In contrast to D_1-CB_1-KO mice, Dlx-CB_1-KO mice (n=8) displayed no impairment compared to wildtype littermates (n=6) in attaining sustained expression of R1. We suggest this may because the putative imbalance created between direct and indirect pathway strengths in D_1-CB_1-KO (the direct pathway but not the indirect pathway expresses D_1) would be cancelled out in Dlx-CB_1-KO, wherein CB_1 are deleted from both the direct and the indirect pathways.



Supplementary figure S. 4 - Aged animals rigidly exploratory despite R1 acquisition. All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. (A) In terms of alternation by surface, the aged R1 trained population (n=11) did not significantly depart from chance level until session 14. (B) The aged R1 trained population showed a similar trend to young adult R1 trained mice to have higher decision latencies than controls, and higher S1 (i.e. exploitatory) than S0 (i.e. exploratory) choice decision latencies, though none of these trends reached statistical significance. (C) Having isolated the 3 aged R1 trained animals who did not reach criterion R1 performance, even with 3 additional training sessions, we checked to see if these 3 animals were also impaired in R1 acquisition. This was not the case: beginning from session 4, this subgroup displayed robustly higher S0 run times than S1 run times, indicating that they had indeed acquired the S1-reward association.

* * *

PART II

* * *

"Got his rag out that evening on the bowling green because I sailed inside him. Pure fluke of mine: the bias."

James Joyce, *Ulysses* (1922).

Investigating hallmarks of 'myside' confirmation bias in a novel mouse model of everyday-like rule revision.

Christopher Stevens, Cathy Lacroix, Mathilde Bouchet, Faustine Roudier, Yamuna Mariani, Giovanni Marsicano, Aline Marighetto.

Abstract

As the daily flow of information available to us increases in speed, complexity, detail, and sheer quantity, our capacity to rationally filter and then integrate it by revising our prior beliefs has become a central question for a wide range of experimental and human sciences. Confirmation, or "myside" bias – over-valuation of novel information which confirms previously internalized cognitive content (beliefs, rules, etc.) and corresponding under-valuation of novel information which disconfirms this same cognitive content – is a serious obstacle to our ability to adaptively revise our beliefs in this complex epistemic environment. Indeed, in modern times, confirmation bias has become a particularly pernicious fact of society, contributing to the propagation of fake news, to the polarization of society, and even to the current scientific replication crisis. Nevertheless, the presence of confirmation bias-like behaviors in non-human animals has never been explored, and therefore little is understood about either its neurophysiological underpinnings or its evolution. In order to advance research in this direction, we designed a novel mouse model of rule revision in such a way that the model environment would be susceptible to elicit myside confirmation bias-like behavior in mice. In the present study, we validate this model and provide the first description of myside confirmation bias-like behaviors in a non-human animal. We also present the results of preliminary investigations using transgenic and aged mice, which enable us to begin disentangling how multiple memory systems contribute to the expression and suppression of confirmation bias.

Introduction

As the daily flow of information available to us, and, to a greater and greater extent, socially imposed upon us, increases in speed, complexity, detail, and sheer quantity, our capacity to rationally filter and then integrate it by revising our prior beliefs finds itself at the center of most critique and analysis of our modern, digital society. Confirmation bias (Nickerson, 1998), specifically in the form now commonly referred to as "myside" bias (Mercier & Sperber, 2017; K. Stanovich et al., 2013; K. Stanovich & West, 2007), has emerged in this environment as a particularly pernicious fact of human cognition, accused of contributing to the propagation of fake news (Lazer et al., 2018), to the polarization of society into camps between which dialogue becomes next to impossible (Del Vicario et al., 2017; K. E. Stanovich, 2021), and even to science specific challenges such as the replication crisis (Baker, 2016; Nuzzo, 2015). The cognitive mechanism of action of myside confirmation bias can be summed up as follows: over-valuation of novel information which confirms previously internalized cognitive content (beliefs, behavioral rules, etc.) and corresponding under-valuation of novel information which disconfirms and calls into question this same cognitive content (Nickerson, 1998). The result is a biasing of all novel information towards the epistemic positions already held by the agent, hence *my-side* bias. As such, its reach goes far beyond the short list of examples above, implicating it in almost every aspect of human interaction involving the updating or revision of beliefs in light of new information: a ubiquitous cognitive phenomenon, identified in some form since antiquity at least (Bacon, 1620; Nickerson, 1998), and hypothesized by some to be an essential, core dimension of how humans reason (Mercier & Sperber, 2017). So, while myside confirmation bias may be historically central to who we are as an evolved species, in today's world, as we are implicitly and sometimes explicitly requested to form opinions and revise our beliefs on an ever-broadening range of increasingly complex and technical topics and events, many of which are literally matters of life and death (COVID response, climate change, etc.), it does seem that this particular cognitive bias is reaching critical pressure.

This points to an important but relatively over-looked dimension of bias, which we can better grasp by tracing its modern English usage back to its roots in the game of lawn bowls. To quote a bowl manufacturer's guide on the subject of bias, "*The behaviour of*

a bowl in play is determined by the weight, velocity of delivery, its bias (or displacement of the centre of gravity from the bowl's centre) and finally, by the nature of the surface upon which it is delivered," (Taylor-Rolph Co. Ltd., 1938). Thus, the bias of a bowl refers to the fact of its center of gravity being displaced; it is inherent to the object. However, the significance of this inherent bias emerges only during play, in interaction with the bowling lawn. In order to have the bowl come to rest at a desired spot, one must compensate for the inherent bias by delivering it towards a point some distance orthogonal to the target. And an inexpert player, not sufficiently aware of this inherent bias, will most likely deliver the bowl wide of their intended mark. Nevertheless, as the scene evoked in James Joyce's *Ulysses* illustrates, bias may also, by a stroke of pure luck or "fluke", transform a flawed delivery into an apparently accurate one. In an important sense, this etymological source signposts everything that we need to investigate in order to understand how confirmation bias impacts behavior: 1) By what means the center of gravity of cognition is displaced in the brain (inherent bias); 2) an accurate description of the environment this biased cognition must navigate, and; 3) the level of foreknowledge cognitive agents have about their own inherent bias and how this can be used to compensate for it.

Perhaps surprisingly, it is the first of these, relating to the neurophysiological mechanisms underpinning confirmation bias, that we know the least about (Kaplan et al., 2016). One overlooked reason for this lack of understanding is the absence of *ad hoc* models designed to investigate whether processes like confirmation bias occur in non-human animals, the existence of which would allow for neurobiological investigation of the underpinning mechanisms. Thus, we designed a novel mouse model of rule revision, specifically in such a way that the model environment would be susceptible to elicit myside confirmation bias-like behavior, on condition that the cognitive mechanisms for doing so are indeed present in the mouse brain. Our experimental design therefore leaned on this twofold assumption: that the observable consequences by which we identify myside bias in humans primarily result from particularities of the (epistemic) environments our species has created for itself to navigate, and therefore there is no empirical reason to suppose that the inherent neurobiological component of myside bias (the bias *in the bowl*) is unique to humans, as has recently been claimed (Mercier & Sperber, 2017). Indeed, in terms of accurately understanding the underpinning neurophysiology of any cognitive phenomenon, lines of

investigation must encompass the direct and experimental study of its evolution (Cisek & Hayden, 2022).

Our model consists of two steps; a rule training phase followed by a rule revision phase. Note that what we refer to here as behavioral "rule revision" should be understood synonymously with "belief revision", insofar as both beliefs and rules constitute policies for governing action in given states, i.e. what are referred to as "state-action policies" in the language of computational reinforcement learning (Sutton & Barto, 2018). Across two papers, of which this is the second, we present a detailed description of both experimental phases, along with fine-grained analysis of the behaviors observed in young adult C57Bl6/J mice as well as transgenic and aged mice. To briefly recap the first phase (Stevens et al., 2022a), we subjected mice to a learning schedule in the 8-arm radial maze which we qualified as a protocol of "indoctrination", broadly defined as any educative process through which natural exploratory drives are inhibited. Our goal in that phase was to robustly impart an initial tactile stimulus-response (S-R) behavioral rule (R1), whereby mice would learn to choose between two surfaces, only one of which was ever predictive of reward location (S1), while the other surface always predicted absence of reward (S0). We furthermore demonstrated that sustained R1 expression, but not R1 acquisition, relies on modulatory inhibitory control over the dorso-striatal direct pathway, the mechanism through which the innate exploratory drive is inhibited. In the second phase, presented here, R1 trained mice were introduced to a new, highly novel radial maze environment where, in order to perform a hippocampus dependent everyday-like memory (EdM) task based around the spontaneous rodent behavior of spatial alternation (Al Abed et al., 2016), they had to simultaneously revise the previously learned R1 rule. The result is a task we refer to as everyday-like rule revision (EdRR). Specifically, the EdRR task was designed in such a way that, as mice perform it, the outcomes from their trial-by-trial performances generate a sequence of confirmations and disconfirmations of the R1 rule. We accomplished this by adding the two surfaces from the first phase, S1 and S0, to the classical EdM environment such that every EdM trial equally constituted an instance of choice between S1 and S0. Crucially, however, S1 was no longer a stimulus predictive of reward location, but merely one among other incidental spatial details of the environment. Behind this design was the desire to more accurately model real world, differentially complex and ambiguous situations of belief revision, in contrast to classical "pure" reversal protocols incapable

of generating sequences of confirmations and disconfirmations, and thereby also incapable of reproducing phenomena such as confirmation bias. Referring back to our analogy, our full model can therefore be understood as a first phase in which we shape the inherent bias (i.e. we train mice to the point of internalizing the R1 rule, thereby creating the "myside" analog), and a second phase in which we deliver the now biased animals into a novel environment (in which they are obliged to revise R1) and observe to what extent their behavior recapitulates human myside bias-like phenomena.

Our observations during revision of R1 reflect myside confirmation bias in several important ways. The vast majority of errors committed in the EdRR task were biased towards S1. However, no major impact on overall number of EdRR errors was observed when compared to controls, mirroring the now well-established finding in humans that strength of myside bias is independent of general intelligence (K. Stanovich et al., 2013; K. Stanovich & West, 2007). Initial exploratory sampling of the novel environment was not affected, indicating that the primary cognitive effect of previous learning was not a reduction of curiosity towards novelty per se but rather a dysfunction of evaluation and integration of the novel environmental information, as per our definition of myside bias above. Indeed, we observed an important lag, over repeated EdRR sessions, in the updating of the likelihood of choosing S1, despite R1 being disconfirmed significantly more than it was being confirmed in the EdRR environment. The probability of choosing S1 was also strongly correlated to EdRR trial complexity, meaning that with more cortico-hippocampal mnemonic uncertainty came higher likelihood of reverting to the striatal R1 response strategy. Analyzing more fine-grained behavioral measures, despite S1 choices consistently leading to more errors, we observed signs of persistent and "irrational" post-choice over-confidence in S1 responses. Moreover, on trials where animals physically revised their initial choice, this revision occurred significantly more often towards a final S1 choice. Based on preliminary investigations using transgenic and aged mice, we suggest that R1 bias and its gradual inhibition are largely, but not wholly, independent of the capacity to perform the EdRR task and therefore of the related cortico-hippocampal functions, which is once again commensurate with the lack of correlation observed between myside bias and general intelligence in humans. Finally, we frame our findings in the context of a multiple memory systems (McDonald & Hong, 2004; McDonald & White, 1993; White & McDonald, 2002) interpretation of everyday cognitive function and show how tasks like EdM and EdRR are uniquely suitable for

investigating the delicate balance between the contributions of such multiple systems and how it can be affected by cognitive state, mental disorder, and ultimately ageing.

Materials & Methods

Animals: Young (8 to 12 weeks) C57BL/6J male mice were obtained from Charles River and collectively housed in a standardized animal room (23 °C; lights on 7 AM to 7 PM; four or five mice per cage). Mice from the aged cohort (18 months) underwent ageing in collective housing on site at the animal facility of the Neurocentre Magendie. D_1-CB_1-KO mice were generated as previously described (Monory et al., 2007; Terzian et al., 2011) by crossing CB_1 floxed mice (Marsicano et al., 2003) with D_1-Cre line mice (Lemberger et al., 2007), in which the Cre recombinase was placed under the control of the D_1 gene (Drd1a). As previously described (Zerucha et al., 2000; Monory et al., 2006), Dlx5/6-Cre mice were crossed with $CB_{1}^{f/f}$ mice to obtain $CB_{1}^{f/f;Dlx5/6\text{-}Cre}$ (here called Dlx-CB_1-KO) and their $CB_{1}^{f/f}$ (WT) littermate controls. Eight to 14-week-old naive male D_1-CB_1-KO and Dlx-CB_1-KO and respective WT littermates were used. All animals were moved to individual cages 2 weeks before the beginning of experiments.

Food restriction: Five days prior to the first day of training, all animals were placed under a progressive food restriction schedule in order to gradually bring them to 85% to 90% of their baseline free feeding weight. Individual animal weight and welfare was monitored daily throughout the duration of the experimentation. All experiments were conducted in accordance with European Directive 2010-63-EU and with approval from the Bordeaux University Animal Care and Use Committee CCEA50. All efforts were made to minimize suffering and reduce the number of animals used.

Viruses and surgery: D_1-CB_1-KO mice were anesthetized in an induction box containing 5% Isoflurane (Virbac, France) before being secured in a stereotaxic frame (Model 900, Kopf instruments, CA, USA) in which 1.0% to 1.5% isoflurane was continuously supplied via an anesthetic mask for the duration of the surgery. Animals were injected with local analgesic (Lidocaine/Lidor, 2mg/ml, 100ul per mouse) and opioid analgesic (Buprenorphine/Buprecare, 0.3 mg/ml,100 ul per mouse) at the beginning of surgery. For viral intra-striatal AAV delivery, AAV vectors were injected with the help of a microsyringe (0.25 mL Hamilton syringe with a 30-gauge beveled needle) attached to a

pump (UMP3-1, World Precision Instruments, FL, USA). D_1-CB_1-KO mice were injected directly into the striatum (STR) (1 µl per injection site at a rate of 0.5 µl per min, for a total of 8 µl per animal), with the following bilateral coordinates: AP 1.5; ML ± 2; DV –3.5 / –3, and AP –0.5; ML ± 2.6; DV –3.5 / –3. Following virus delivery at each site, the syringe was left in place for 2 minutes (DV –3.5 sites) and 5 minutes (DV –3 sites) before being slowly withdrawn from the brain. 6 mice were injected with pAAV-CAG-flexx-IRES-mCYT (empty control vector) to create the D_1-CB_1-KO-$Str^{-/-}$ group and 6 mice with pAM-CAG-flexx-CB1myc to induce re-expression of the CB_1 receptor gene in the striatum and create the D_1-CB_1-KO-$Str^{+/+}$ group. At the end of surgery, all operated animals were given an anti-inflammatory injection (METACAM, 2mg/ml, 50ul per mouse ip.). In this experiment, expression was allowed to take place for 4-5 weeks after local infusions. Mice were monitored and weighed daily post-operation for three days and also given one more daily i.p. injection of Metacam, as described above. All animals regained their pre-surgery body weight, meaning none were excluded from the experiments.

Radial maze: The behavioral apparatuses used are 8-armed fully automated radial mazes (Imetronic), the surface of which is raised ~100cm off ground level. Access to each arm is from a central platform by means of automated vertically retracting doors. When all doors are closed during behavior, the experimental animal is contained within the central platform, a regular octagon of size ~485cm^2 and edge 10cm (i.e. the width of each arm and door). At the distal end of each 50cm length arm is an automated pellet distributor for dispensing food reward. The distributor is set into a slight indent in order to hide its state (i.e. baited or not baited) from the animal. For this study, we produced removable polymer panels which could be placed so as to cover the entire area of the radial maze. The panels, all painted the same color as the radial maze, were of two distinct tactile finishes: smooth surfaced panels (similar to the usual surface of the radial maze) and irregular surfaced panels (the finish of which was a uniform but irregular beveled pattern of < 2mm maximum relief). This allowed us to present the radial maze according to various tactile configurations: entirely smooth, entirely irregular, or various combinations of smooth and irregular. Animal movements in the radial maze are detected via video camera and motion detection software (GenCam) using either visible or infra-red light, depending on the experimental conditions (see below). The motion detection software communicates with a second piece of software, POLYRadial

(Imetronic), through which pre-programmed sequences of automated radial maze actions are triggered. This program is used for the design and execution of behavioral exercises (sequences of door openings, location of food reward, conditions for opening and closing of doors, etc.). Hence, the exercises are customizable and contingent upon a combination of both the detected movements of the animal and automated timed sequences.

Behavior

Habituation: Prior to the first day of tactile discrimination learning, all animals were habituated to the context and functioning of the radial maze apparatus. Food restriction, as described above, began three days before habituation (i.e. five days before training). At the beginning of each habituation session, the animal was placed by the experimenter in the central platform of the radial maze, all 8 doors of which were closed. Once removed to the control room, the experimenter launched the habituation program via the POLYRadial software. The habituation program began by an interval of 10 seconds during which the animal could explore the central platform. Following this, all 8 doors opened simultaneously, presenting the animal with the opportunity to freely explore the entire surface of the maze. As the animal explored, once it had advanced to the most distal section of a given arm (location of the distributor and food reward) and returned to the central platform, the door of that arm automatically closed behind it, thus preventing further access to that arm in the current session. Thus, once the animal had fully explored all 8-arms, it found itself again contained within the central platform. At this point, a further habituation session could be launched if needed. It was considered that when an animal had recovered and consumed at least 5 out of 8 available food rewards in a single session that it was fully habituated to the relevant functionalities of the apparatus. All animals reached this habituation criterion within an average of 5 sessions. Since the tactile discrimination phase is conducted in the dark, thus potentially making it difficult for animals to perceive that doors always opened in contiguous pairs, animals also underwent a second habituation session 24 hours after the first, during which pairs of doors opened simultaneously, creating a choice to explore one of two neighboring arms, as would be the case in the subsequent tactile discrimination phase. Crucially however, during both phases of habituation, the surface of the radial maze

was entirely covered in the surface type (smooth or irregular) to which a given animal was to be assigned during the subsequent tactile discrimination training. In this way, even during habituation, animals assigned to learn to associate, for example, the irregular surface with reward location had no prior experience of being rewarded on the smooth surface, and vice versa.

Tactile discrimination: In a preliminary phase (figure 1a), mice were trained in the radial maze to discriminate between two surface types, one smooth and one irregular, in a stimulus-response (S-R) manner. For each animal, only one of these surface types was predictive of reward location (S1) while the other surface was predictive of no-reward (S0). This preliminary task was conducted in darkness in order to remove the capacity for visual spatial orientation, thereby obliging reliance on other sensory inputs, notably tactile. Each trial consisted of choosing between two neighboring arms to visit, one S1, one S0. Mice were considered to have fully internalized this S1-reward association rule (R1) when they had reached a mean performance of 75% correct responses averaged across two sessions.

In order to successfully express R1, animals had to inhibit and overcome their innate exploratory drive to explore both surfaces equally and instead adopt a "win-stay" strategy (McDonald & White, 1993) with respect to S1. During this phase, ethanol (70%) was also used to add a specific odor-based dimension to the R1 environment. For full description and behavioral analysis, see (Stevens et al., 2022a).

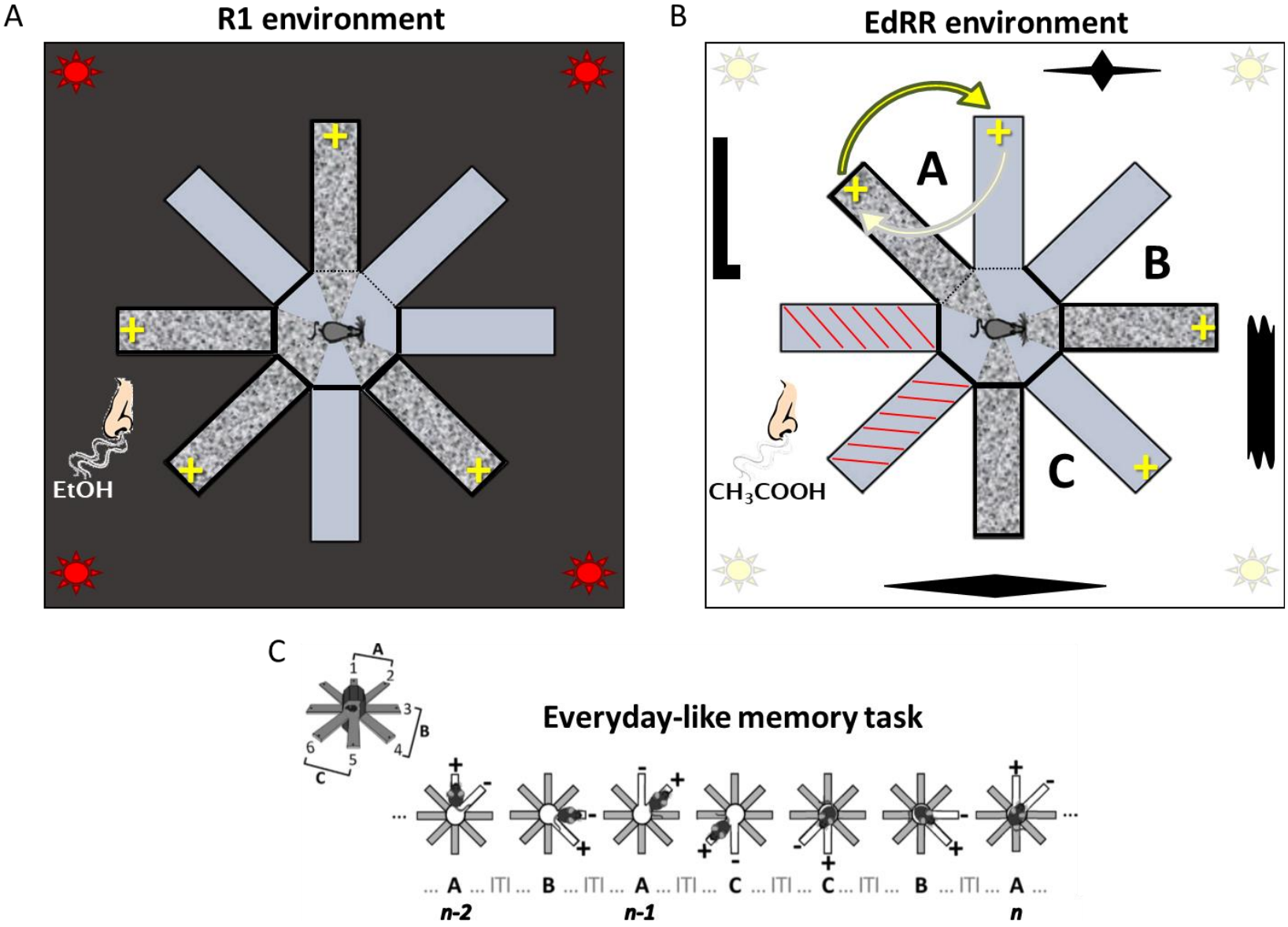

A
R1 environment
EtOH
B
EdRR environment
A
B
C
L
CH3COOH
C
1
2
3
4
5
6
A
B
C
Everyday-like memory task
... A ... ITI ... B ... ITI ... A ... ITI ... C ... ITI ... C ... ITI ... B ... ITI ... A ...
n-2
n-1
n

Figure 6 - Behavioral environments and tasks.
Graphical representation of the radial maze environments composing the full everyday-like rule revision protocol. (**A**) The R1 tactile discrimination phase is conducted in darkness, removing all visual spatial bearings, with the objective of focusing attention on the tactile dimension. The surface of the radial maze is divided in two according to surface: a smooth surface, represented in solid gray, and an irregular surface, represented in dappled gray. Each animal learns that one of the two surfaces is always predictive of food reward location (yellow '+' symbols). This surface is referred to as S1. Whether S1 is the smooth or the irregular surface is counter-balanced among the animals. All 8 arms are used in this phase. Each trial consists of a choice between two contiguous arms, one S1, one S0. A 70% ethanol mix is also used to give this R1 environment a distinct olfactory dimension. (**B**) Once animals have completed R1 training (either by reaching 75% criterion level averaged across two sessions or by completing a fixed number of R1 sessions, depending on experimental conditions; see main text) they are moved to the EdRR radial maze environment, which contains several novel differences with respect to the R1 environment: radial maze situated in a separate room; lights on, hence visibility of spatial landmarks (represented by the black shapes around the radial maze); 1% acetic acid mix used to distinguish environment at the olfactory level. The EdRR task, similar to the classical EdM task, uses 6 arms of the radial maze, arranged into three pairs, A, B, and C. However, EdRR is distinct from EdM by the presence in each pair of one S1 and one S0 arm. The cognitive consequence of this is that every EdM trial (spatially alternate relative to previous choice on present pair; see below) simultaneously constitutes an instance of an R1 trial (choose S1 rather than S0). This allows us to qualify each EdM error made during performance of EdRR as either an R1/S1 or a nonR1/S0 type error, providing the basis for the quantification and analysis of R1 bias during everyday-like revision of R1. (**C**) Illustrative sample sequence of classical EdM task. On trial *n-2*, the animal chose the left arm of pair A and was rewarded. Trial *n-1* on pair A then constitutes a complexity level 1 trial, as there has been 1 interposed trial on pair B between *n-2* and *n-1*. The animal correctly chooses the right arm on *n-1*. When pair A is next presented, it constitutes a complexity level 3 trial, as there have been 3 interposed trials (pairs C, C, and B) between *n-1* and *n*. To choose correctly on trial *n*, the animal must focus on recalling which arm it chose on trial *n-1*, inhibiting interference from both the interposed trials on other pairs and from the memory of trial *n-2*.

Everyday-like memory: As previously described (Al Abed et al., 2016), the aim of the everyday-like memory (EdM) task in the radial maze is to model daily life situations in which numerous, repetitive events, of varying complexity and often with only subtle but important changes in their content must be accurately remembered and recalled. For example, we have to keep in memory multiple pieces of constantly changing information such as "Where did I park my car?", "Did I feed the dog?", "What groceries need replacing?", etc., for variable periods, all while occupied with other daily activities, until that information is needed again and mobilized. These situations require both mnemonic retention and continuous inhibitory organization/updating of memories as they are used, in order to avoid interference from other similar memories. The demands on mnemonic retention and inhibitory organization are oppositely proportional with respect to the temporal interval between repetitions. To illustrate: the more frequently I park my car in the same zone, the lower the demand on mnemonic retention (i.e. "Where did I park at work this morning?" versus "Where did I park at the airport two weeks ago?"), but the higher the demand on inhibitory mnemonic organization in order not to mix up successive placements ("No! This is where I parked *yesterday* morning!").

In the EdM model (figure 1b), animals must simultaneously (i) store items of information relative to three distinct task contexts (arm pairs; A, B, and C), i.e. which was the most recently visited arm from each of the three sequentially presented arm pairs (figure 1c), and (ii) use and update each of these items of stored information in order to correctly choose which arm to visit out of the currently presented pair. Mice have no way of predicting which of the three task contexts/arm pairs they will be presented with on any given trial, and thus no way of knowing, while waiting in the central platform for the next trial to begin, which memory they will next need to recall nor which ones they will need to inhibit. Identification of which pair is which depends on spatially identifying its position relative to prominent extra-maze cues in the experimental room (represented by the black shapes surrounding the maze in figure 1b).

In each session of EdM, animals are presented with a sequence of 23 trials, each trial consisting of one presentation of one of the 3 arm pairs (A, B, and C). This sequence changes from session to session, and is pseudo-random (i.e. unpredictable) from the mouse subject's perspective. Mice must choose to visit one arm out of an arm pair in each trial. On a given trial, the reward will always be located in that arm which was not

visited by the mouse on the previous presentation of the same arm pair, whether their previous choice was correct or incorrect. In other words, the reward in a given pair switches arm only once it has been retrieved (compare what happens with pairs A and C to what happens with pair B in the sample sequence represented in figure 1c). The task therefore implies retaining in memory which arm was visited on any given arm pair presentation *n* (the "sampling" trial) until the next trial consisting of a presentation of the same arm pair, *n+1* (the "testing" trial). In short, the EdM task relies on and reinforces the spontaneous mouse behavior of spatial alternation and, in contrast to the "win-stay" tactile discrimination R1 rule, requires a "win-shift" strategy (McDonald & White, 1993). The number of interposed trials on either of the other two arm pairs, plus the duration of the inter-trial intervals (ITI), constitute the retention component of the task. An organizational mnemonic component is also present in both the necessity to inhibit pair-specific interference on any given arm pair presentation *n*, so that the previous *n-1* choice and not the previous-previous *n-2* choice is recalled (figure 1c), and also in the need to inhibit mnemonic content relative to the other two pairs which would constitute intrusive and interfering cognitive noise to the recall process specific to the present pair.

At the beginning of each EdM session, the animal is placed at the center of the maze, with all vertically opening arm access doors in the up/closed position. After a 10 second pause, the two doors to one of the 3 pairs (A, B or C) open simultaneously by retracting below the surface of the maze, whereupon the mouse can choose to visit the distal, reward zone of one of the two arms. The door of the non-chosen arm closes only once the mouse has reached the reward zone of its chosen arm. When the mouse returns and is again detected in the central platform of the maze, the door of the chosen arm also closes behind it. After a given ITI (in the present study ITIs = 5-10s unless otherwise stated), during which the mouse is confined to the central platform, another trial begins with the opening of two doors, either of the same arm pair, or one of the other two arm pairs, and so on. In the classical version of the EdM task, on the initial per session trial of each pair both arms contain a food reward. This trial establishes how the reward will spatially alternate in the subsequent trials and is not included in the calculation of EdM performance.

The mnemonic challenge of EdM on which we primarily focus in the present study is the *retention difficulty* of each trial, determined by the number of trials on the other two arm pairs interposed between a presentation *n* and subsequent presentation *n+1* of the same arm pair. This gives rise to 5 levels of complexity, denoted by the number of interposed trials on the other two arm pairs, i.e. from 0 to 4. Level 0 therefore corresponds to two immediately consecutive presentations of the same arm pair, with no interposed trials on other pairs, and is thus most similar to a classical T- or Y-maze spontaneous alternation trial (figure 1c; the second presentation of pair C is therefore a level 0 trial; the second presentation of pair A, a level 1 trial; the third presentation of pair A, a level 3 trial, etc.).

The sequences of pairs A, B and C within a training session of the EdM task have been designed such that three consecutive sessions are required in order to equally balance the number of trials of each difficulty level. Hence, behavioral parameters expressed as a function of trial complexity are calculated as the mean or median per block of 3 sessions and analyzed on this basis. Each block of 3 sessions consists of 12 trials of each complexity level, for a total of 60 trials, not including the initial "sampling" trial on each of the three arm pairs which do not figure in evaluation of the EdM performance.

Rule revision: The rule revision modification to the EdM task (everyday-like rule revision, EdRR) consists in its combination with the S1 and S0 surfaces from the tactile discrimination phase (Fig.1). Thus, the primary specificity of the EdRR task consists simply in the presence, on each arm pair and in each trial, of a S1 arm and a S0 arm, counter-balanced left and right between pairs. Also, in contrast to the classical EdM task, in EdRR only the S1 arm is rewarded in the initial trial of each pair in each session, following which reward location spatially alternates as a function of being recuperated. This tactile modification of the EdM environment entails that, as an animal performs the EdRR task, a sequence of confirmations and disconfirmations of R1 is generated: whenever a S1 arm is rewarded or a S0 arm is not rewarded, R1 is confirmed. Conversely, whenever a S0 arm is rewarded or a S1 arm is not rewarded, R1 is disconfirmed. It is on this basis that we present this rule revision model also as a potential model of myside confirmation bias.

Thus, once animals have reached criterion level performance in the R1 environment, they are moved to a new radial maze in a new room, in presence of a new odor (acetic

acid, 1%, as opposed to ethanol), with new lighting conditions (lights turned on) rendering intra- and extra-maze cues visible. The rationale for these contextual changes was to emphatically announce, via multiple sensory systems, that a novel environment, distinct from the R1 environment, has been entered.

In brief, as R1 trained mice perform the EdRR task, they must not only retain and organize their previous actions as per the classical EdM task, but additionally adaptively "forget" the R1 rule in order to ensure that they respond to each trial as an instance of EdM and not as an instance of R1, which would otherwise lead them to repeatedly choosing S1. This addition of S1 and S0 to the EdM environment enabled us to qualify each EdRR choice and error an animal made. We could then calculate both the total number of S1 choices and the proportion of S1 errors to total S1 + S0 errors, providing us with two strong measure for assessing the magnitude and evolution of R1 interference in EdM behavior.

Analysis

All raw data extraction, analysis, statistical comparison, and graphical representation was generated using custom codes written in Python (Van Rossum & Drake, 2009) using the pandas (Reback et al., 2020), numpy, pingouin (Vallat, 2018), bioinfokit (Bedre, 2021), matplotlib (Hunter, 2007), and seaborn (Waskom, 2021) libraries. All code is open source and available at https://github.com/metaphysiology. Here we give brief details about some of the cognitive behavioral parameters we analyzed.

Decision latency: The time taken by each animal between the instant when a trial began (doors of the current trial pair open) and the instant when the threshold from the central platform into the surface-arm of the animal's definitive choice was first crossed (decision latency, milliseconds). We were further able to analyze these decision latencies according to various factors such as surface type of the definitively chosen arm, etc.

Run time: The time taken for animals to travel the distance from the threshold of the definitively chosen arm to the reward-distributor containing distal extremity (run time, milliseconds). As above, we could then analyze this measure according to whether the definitive choice was S1 or S0, etc.

Choice revision: During certain trials, animals crossed the threshold into one arm of a pair but, prior to entering its most distal zone (which would trigger the door to the unchosen arm to close), revised their choice by exiting the arm. At this point animals could either choose to explore the other arm of the pair or (more rarely) re-enter the same arm. As long as the distal zone of either arm had not been entered, this process could technically continue indefinitely. We developed a novel analysis to quantify this behavior, which we took to be an occasional external and physical manifestation of the ongoing cognitive decision-making process. On a given trial, each additional crossing of either of the two central platform-to-arm thresholds, in the direction from the platform towards the arm only, was quantified as one choice revision. Each choice revision was quantified as a 'KOOK' unit, capturing the fact that some choice revisions were ultimately error-inducing, 'KO', while others were rectifying, 'OK'. For detailed discussion of the run time and choice revision parameters, see (Stevens et al., 2022a).

Results

Impact on EdM performance of a previously acquired, biasing cognitive rule.

To begin, we present the results from young C57Bl6/J animals who completed 12 sessions of the everyday-like memory R1 rule revision task (EdRR). This number of sessions allowed us to characterize not only the initial impact but also the evolution over time/repeated training of the previously learned, partially EdM antagonistic R1 rule. We defined the following two groups from three iterations of the experiment; animals who had acquired and expressed the R1 tactile discrimination rule up to criterion level in the first environment (R1 trained population, n = 40), and animals who had been rewarded on every trial in the first environment, regardless of their surface choices (controls, n = 19).

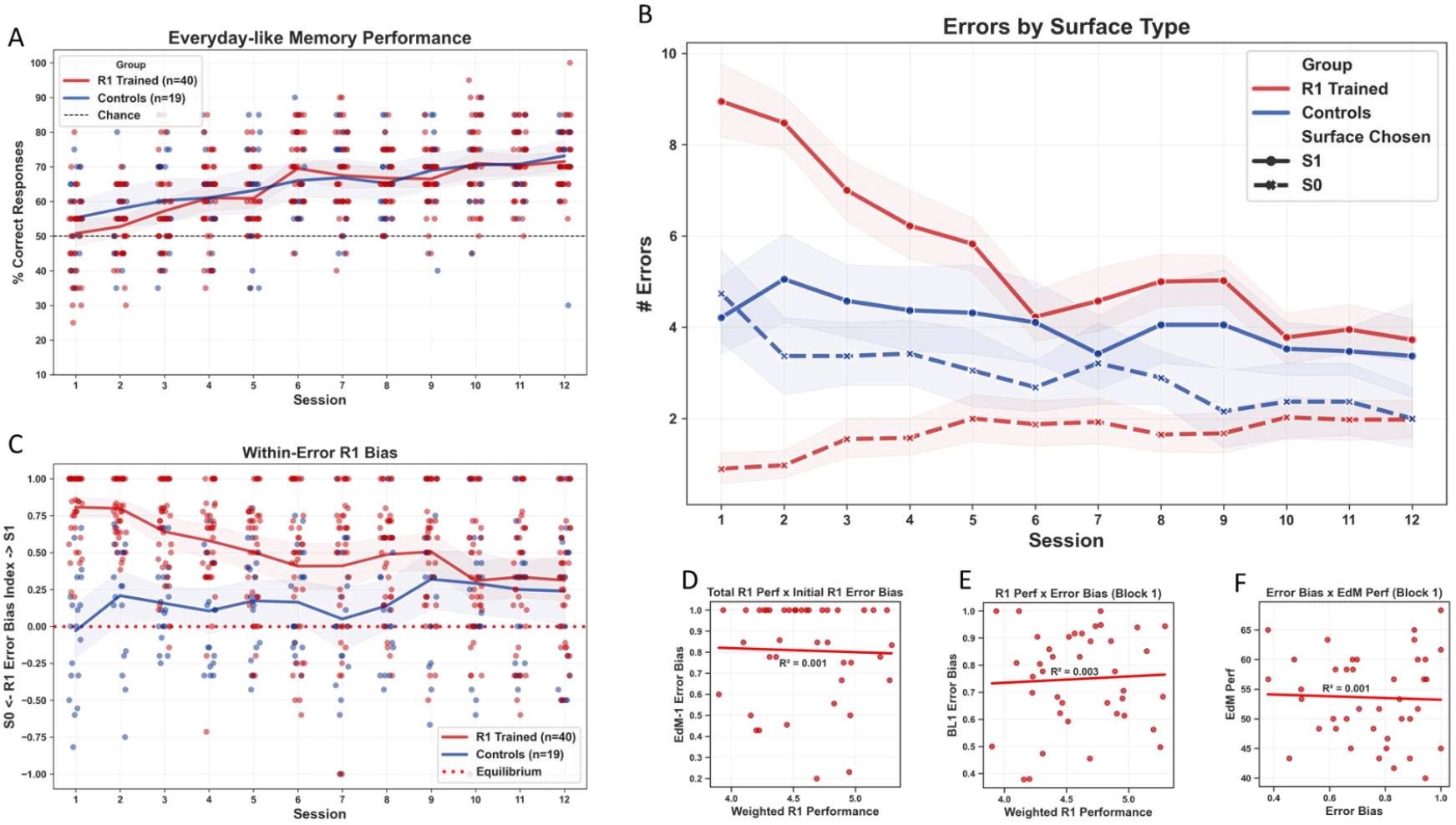

A
Everyday-like Memory Performance
Group
R1 Trained (n=40)
Controls (n=19)
Chance
% Correct Responses
Session
B
Errors by Surface Type
Group
R1 Trained
Controls
Surface Chosen
S1
S0
Errors
Session
C
Within-Error R1 Bias
S0 <- R1 Error Bias Index -> S1
R1 Trained (n=40)
Controls (n=19)
Equilibrium
Session
D
Total R1 Perf x Initial R1 Error Bias
EdM-1 Error Bias
R² = 0.001
Weighted R1 Performance
E
R1 Perf x Error Bias (Block 1)
BL1 Error Bias
R² = 0.003
Weighted R1 Performance
F
Error Bias x EdM Perf (Block 1)
EdM Perf
R² = 0.001
Error Bias

Figure 7 - Impact on everyday-like memory of a previously learned rule especially visible in nature of errors committed.

Experimental animals are represented in red (n=40), controls in blue (n=19). All error bands represent 95% confidence intervals, vertical spaces between bands provide visual indication of statistical significance; detailed statistical analysis in main text. Curves represent population means, dots are individual values. (**A**) Everyday-like memory performance calculated from 20 per session trials. R1 trained and control populations both began with very low, barely above chance level overall performances which improved gradually with repeated training and at a comparable rate. Controls performed slightly but significantly better than R1 trained during first block of 3 sessions only. (**B**) Although both populations made similar overall numbers of EdM errors, during EdRR 1 + 2 especially, ~90% of R1 trained errors were biased towards S1. This proportion decreased with repeated EdRR training, primarily as a function of decreasing numbers of S1 errors. Even in final EdRR sessions, however, significantly more errors occurred on S1 compared to S0. Control animals also displayed a slight emergent bias towards S1, significant when averaged across all sessions but not in any given session (analyzed further below). (**C**) We then calculated a within-error R1 bias index per animal, per session by taking the relative difference between S1 and S0 errors (S1 errors – S0 errors / Total errors) and tracking its evolution. No significant decrease was observed between EdRR 1 and EdRR 2 in R1 trained animals, after which the within-error R1 bias began to gradually decrease, though even after 12 sessions of EdRR errors were still biased towards S1. Controls also developed a slight S1 error bias over repeated EdRR training, which achieved statistical significance in the final 4 sessions. (**D**) + (**E**) Mirroring human studies showing no correlation between strength of myside bias and overall cognitive ability, linear regression analyses revealed no correlation between strength of individual performance levels in the R1 training phase and either first EdRR session or first EdRR block (composed of sessions 1-3) levels of within-error R1 bias. (**F**) Linear regression analysis also found no correlation between level of within-error R1 bias and overall EdM performance across the first block, indicating that how biased an animal was in the errors it committed was no predictor of how many errors overall it would make, and vice versa, once again very closely mirroring findings in human studies of myside bias.

To recall briefly, reward location in the classical EdM task is determined by a rule of trial-by-trial spatial alternation between the arms of each given pair (total 3 pairs; A, B, and C, see figure 1b). In EdRR, each trial of the EdM task also constitutes an instance of the R1 choice between visiting an S1 arm or an S0 arm. This implies the necessity for the animal to ignore/inhibit its prior R1 learning in order to perform in a purely EdM manner. If not, it risks making consistent EdM errors by acting according to R1 and thereby over-visiting S1 arms rather than spatially alternating (refer to Materials & Methods for detailed description). In order to contextually mark the transition from the R1 to the EdRR task, we designed the two task environments to be as novel as possible with respect each other; new room and radial maze, change of visibility conditions from darkness to lights on, thus presence of visual spatial landmarks in EdRR, plus a change in ambient odor.

1. Nature – more than number – of EdM errors impacted by previous rule learning.

EdM performance: R1 trained and control animals performed comparable levels of correct responses per session and displayed comparable learning curves over the 12 sessions/4 blocks of training, with a statistically significant difference only when averaged across the first block (sessions 1-3) for the R1 trained population to perform worse than controls (figure 2a; one-way ANOVA with pairwise Tukey post hoc; Block 1, $F(1, 175) = 4.9$, $p = 0.028$). Individual performances (R1 trained population only) in the EdRR task also did not correlate with either averaged (supplementary figure S.1a, $R^2 = 0.002$) or weighted averaged R1 performances (using a discount factor to give older R1 session performances less weight; supplementary figure S.1b, $R^2 = 0.00$).

Within-error R1 bias: Beginning with the first session (EdRR 1), though similar in terms of mean total number of errors (EdRR 1; R1 trained population, mean errors 9.85; controls, mean errors 8.95), the nature of these errors was vastly different between the two groups. During EdRR 1, the mean ratio of errors on S1 arms to S0 arms was around 9:1 in the R1 trained population, whereas in controls this ratio approached 1:1 (4.2:4.7) (figure 2b; mixed ANOVA design, interaction between 'Group' within 'Surface', $F(1, 57) = 47.4$, $p < 0.0001$). Interestingly, during EdRR 2, conducted 24 hours after EdRR 1, this mean error ratio decreased only very slightly in the R1 trained population, to around 8.5:1. We also analyzed and represented this error bias and its evolution over

repeated EdRR training as a relative difference (S1 errors – S0 errors / Total errors), calculating per animal and per session what proportion of its total errors were biased towards the previously learnt R1 behavioral rule. This revealed that in many R1 trained animals, especially in the first EdRR sessions where the greatest total number of errors were made, 100% of these were committed on S1 arms (figure 2c). Overall, the R1 trained population made a significantly higher proportion of S1 errors than controls (figure 2c; one-way ANOVA with pairwise Tukey HSD post hoc, $F(1, 705) = 96$, $p = 0.001$). No significant decrease in this within-error R1 bias was observed between EdRR 1 and EdRR 2. We did also observe a trend in controls to commit a comparatively higher proportion of errors on S1 arms than on S0 arms. This was significant when averaged across all sessions and will be further analyzed below (repeated measures ANOVA, $F(1, 18) = 20.5$, $p < 0.0003$).

These results demonstrate that the strongest impact of prior R1 learning was manifest not in EdM performance itself (although it was initially observable here also) but rather in the nature of the errors committed. Linear regression analyses, however, revealed no correlation between final individual R1 performances and strength of within-error R1 bias in EdRR 1 (supplementary figure S.1c, $R^2 = 0.018$). Neither was a correlation found when we accounted for the full R1 historic of each animal by comparing total weighted R1 performance (using a discount factor to give more weight to later R1 sessions compared to earlier, more temporally distant ones) against the within-error R1 bias in either EdRR 1 (figure 2d, $R^2 = 0.001$) or in block 1 (i.e. EdRRs 1 – 3; figure 2e, $R^2 = 0.01$). Finally, individual within-error R1 bias index values in block 1 also did not correlate with EdRR performance across the same sessions (figure 2f, $R^2 = 0.001$). Taken together, we translate this lack of simple correlations between behaviors as preliminary evidence that the cognitive output elicited by the EdRR task is the result of multiple cognitive systems interacting in a manner too complex to be easily reduced. It also indicates that levels of cognitive ability in each of these systems are, at least to some meaningful degree, independent of each other.

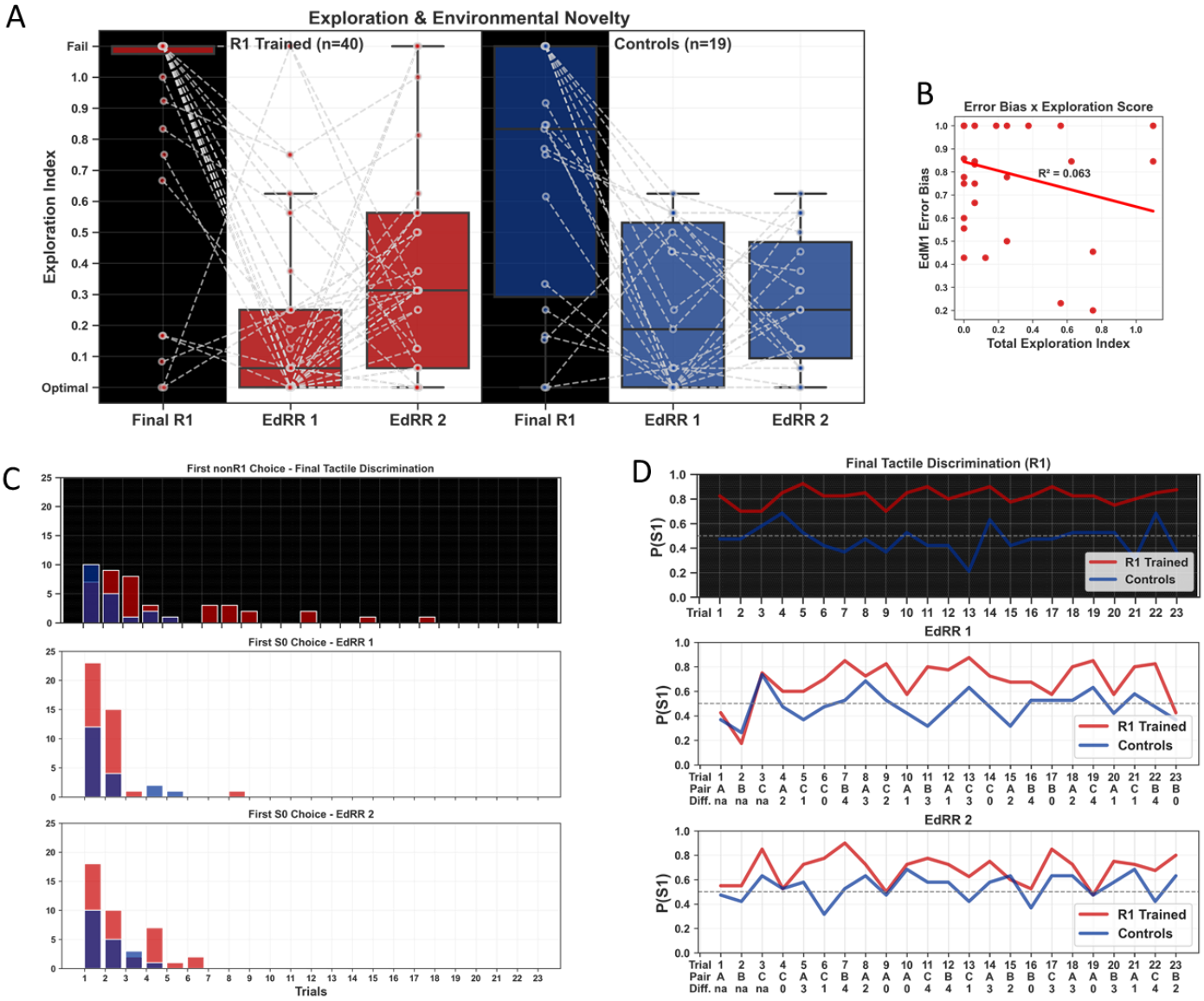

A
Exploration & Environmental Novelty
R1 Trained (n=40)
Controls (n=19)
Exploration Index
Fail
Optimal
Final R1
EdRR 1
EdRR 2
B
Error Bias x Exploration Score
EdM1 Error Bias
Total Exploration Index
R² = 0.063
C
First nonR1 Choice - Final Tactile Discrimination
First S0 Choice - EdRR 1
First S0 Choice - EdRR 2
Trials
D
Final Tactile Discrimination (R1)
EdRR 1
EdRR 2
P(S1)
R1 Trained
Controls
Trial
Pair
Diff.

Figure 8 - High initial environmental novelty-driven exploratory behavior decreases as novelty fades.

Transfer from the R1 to EdRR environment constitutes high multi-sensorial novelty (see figure 1). Here we assess the impact of this novelty on behavior. R1 trained animals (n=40) represented in red, controls (n=19) represented in blue. (**A**) Exploration index (EI) per animal per session, calculated from number of trials (as a function of total number of trials) within which animals explored each available arm at least once. Low EI values indicate all available arms were explored within a low number of trials, high EI values that certain arms were visited repeatedly at the expense of other arms not being explored. Boxplots represent median and interquartile range. Values for final R1 session (black background columns), EdRR 1, and EdRR 2. Certain outliers notwithstanding, the R1 trained population, who explored less than controls in the final R1 session, nevertheless explored the novel environment during EdRR 1 at almost optimal levels, with a lower median population EI even than controls. During EdRR 2, in the same and therefore, logically, now less novel environment, the R1 trained population, but not controls, had a significantly higher median EI score, which we will see corresponded with higher initial probability of choosing S1 ('P(S1)'). (**B**) Linear regression analysis revealed no correlation between individual EI values and strength of within-error R1 bias, indicating independence of the cognitive mechanisms responsible for these behaviors. With respect to this, note in (A) that certain optimal explorers in EdRR 1 failed to fully explore all arms in EdRR 2. (**C**) Analyzing in more detail, we represented as histograms the trials in which animals made their first S0 choice (1 bin = 1 trial). This revealed that R1 trained animals were already choosing earlier than could be predicted from R1 performance levels in the final R1 session, but this leftward skewing increased significantly in EdRR 1 with 57.5% or R1 trained animals choosing S0 on the first trial. This proportion decreased in EdRR 2, though still skewed towards the beginning of the session. Controls first chose S0 according to a population probability closer matching random distribution (see main text for details). (**D**) Average population level P(S1) per trial for final R1, EdRR 1, and EdRR 2. This allows visualization of pronounced initial exploratory behavior in trials 1 + 2 of EdRR 1, followed by persistently high P(S1 values in remaining EdRR 1 trials and in EdRR 2.

2. Environmental novelty and exploratory drive: R1 responding increases as novelty decreases.

Novelty boosts exploration: As we have just seen, the R1 trained population displayed marked R1 error bias in the EdRR task without, however, committing many more total errors than controls, meaning they did also visit S0 arms. Moreover, by analyzing intra-session trial-by-trial behavior, we were able to observe that, in EdRR 1 especially, the primary driver of S0 exploration was not, as may have been expected, an accumulation of unrewarded visits to S1 arms triggering a "lose-shift" strategy response towards S0. Rather, upon first placement in the novel EdRR environment, the R1 trained population explored all available arms, S1 and S0, on average even more methodically and efficiently than controls (figure 3a). To quantify this behavior, we calculated an exploration index (EI) for each animal, with lower values indicating earlier exploration of all 6 available arms and higher values meaning that certain arms had been repeatedly visited at the expense of others not yet visited. The EI was therefore a function of the minimum number of trials required in a given EdRR session trial sequence for all 3 arm-pairs to be presented at least twice. In EdRR 1, for example, this occurred with trial number 7. Animals who had explored all 6 arms by trial number 7 of EdRR 1 were thus attributed an optimal EI of 0, labelled 'Optimal' on the y-axis in figure 3a (optimally exploring animals in EdRR 1; R1 trained, 20/40, 50%; controls, 6/19, 31.6%). Animals who explored all 6 arms at least once by trial 8 were attributed an EI of 0.0625 (R1 trained, 7/40, 17.5%; controls, 3/19, 16%), and so on. Only 2/40 (5%) R1 trained animals failed to explore all 6 arms of the novel radial maze by the end of EdRR 1. These animals were attributed an EI of 1.1, labelled 'Fail' on the y-axis in figure 3a. Overall, the R1 trained population had a median EI of 0.0625 in EdRR 1, i.e. one degree above optimal, whereas the control group had a median EI of 0.1875, though there was no statistically significant effect of 'Group' in EdRR 1 EI values. In brief, the high initial novelty of the EdRR environment seems to have either amplified the exploratory drive or, we suggest instead, inhibited ongoing active inhibition of the exploratory drive (see below and (Stevens et al., 2022a)). Either of these scenarios could explain why the effect of novelty on exploration was actually slightly more prominent in the R1 trained population than in controls, since for R1 trained animals "explore", as we shall see in

more detail below, translates behaviorally as “choose S0”, in opposition to “choose S1” which is the S-R “exploit” command they were trained on in the preliminary R1 phase.

As novelty fades, interference invades: In order to more fully appreciate the role of novelty in this prominently exploratory behavior, antagonistic to R1 persistence, we posited that any effect of novelty should be maximal upon first exposure to the novel environment (i.e. EdRR 1) and therefore, logically, lessened in subsequent sessions. We therefore calculated the EI values from both the previous (final R1) and subsequent (EdRR 2) sessions and compared these to the EI values from EdRR 1. Due to the intentionally accentuated qualitative environmental differences between the EdRR environment and the prior R1 environment, direct within-individual comparison between final R1 and EdRR 1 could at best be indicative (R1 had no spatial landmarks, animals were therefore less able to discern whether they had, for example, visited all S0 arms or simply one S0 arm several times; R1 uses all 8 arms of the maze, EdRR only 6, etc.). However, comparing the final R1 session behavior (black columns in figure 3a) between R1 trained and control animals, we observed that 30/40 R1 trained animals (75%) did not fully explore the R1 environment compared to only 6/19 control animals (32%), leading in turn to a significant difference in overall exploratory performance (one-way ANOVA with pairwise Tukey HSD post-hoc between groups, with unbiased Cohen effect size, $F(1, 57) = 6.12$, $p = 0.016$, $d = 0.69$). Next, as seen above, both groups displayed an increase in exploratory behavior when introduced to the novel EdRR environment. Upon second exposure to the EdRR environment (i.e. EdRR 2), however, the median EI value significantly increased as compared with EdRR 1 in the R1 trained population but not in controls, due to a relatively increased preference for choosing S1 (figure 3a; R1 trained, 0.0625 EdRR 1 vs 0.3125 EdRR 2, paired t-test, $t(39) = 2.6$, $p = 0.013$; controls, 0.1875 EdRR 1 vs 0.25 EdRR 2, paired t-test, $t(18) = 0.3$, $p = 0.76$). The number of individual R1 trained animals who failed to explore all available arms by the end of the session also increased from 2/40 (5%) in EdRR 1 to 5/40 (12.5%) in EdRR 2, whereas in the control group no animals failed to explore all available arms in either EdRR 1 or 2. Yet, strikingly, all R1 trained animals who failed to fully explore during EdRR 2 had been optimal or quasi-optimal explorers in EdRR 1 (EI < 0.1). Moreover, there was no correlation between how quickly R1 trained animals explored all available arms and how biased they were in their errors during the rest of EdRR 1 (figure 3b, $R^2 = 0.063$) or EdRR 2 (supplementary figure S.1d, $R^2 = 0.007$), once again

indicating that differentially powerful and separate, albeit interacting, cognitive mechanisms were responsible for each of these behaviors.

Exploring space and rules: We next identified and compared on which trial each animal made its first S0 choice, again in the final R1 session, in EdRR 1, and in EdRR 2 (figure 3c). Looking first at the final R1 session (figure 3c, top), we can see that even here, despite having learnt that S0 was not associated with reward location, the majority of R1 trained animals made their first S0 (i.e. exploratory) choice towards the beginning of the session (7, 9, and 8 animals out of 40, i.e. 17.5%, 22.5%, and 20% of R1 animals in the first, second, and third trials, respectively). The remaining 16/40 R1 trained animals (40%) made their first S0 choice somewhere between the 4th and 18th trials out of a total of 23 or 24 trials during their final R1 session (the timing of first S0 choice in final R1 session was not predictive of R1 performance; see (Stevens et al., 2022a) for full analysis and discussion of exploratory behavior during R1 training). In the case of control animals, it was expected that they would make their first S0 choice in the R1 environment according to a random choice behavior compatible distribution, and this is what we observed; all control animals first chose S0 within the first five trials according to a distribution of 10, 5, 1, 2, and 1 individuals out of 19 per trial, i.e. 52.6%, 26.3%, 5.2%, 10.5%, and 5.2%, respectively. However, when initially introduced to the novel EdRR environment, the proportion of R1 trained animals choosing the S0 arm on the first trial increased to 23/40 (57.5%). 15 of the remaining 17 R1 trained animals (15/40, 37.5%) chose S0 on the second trial, 1 on the third trial and, finally, the last of the R1 trained animals on the 8th trial (figure 3c, middle). Importantly, the timing of these initial EdRR environment S0 choices can be explained neither by strength of individual R1 performance in the final R1 session 24 hours earlier, nor by purely random choice behavior, again indicating a pointedly exploratory boost primarily due to environmental novelty per se. Recall that the very first presentation of each pair in each session of EdRR necessarily constitutes a confirmation of R1; in these three trials, S1 arm choices are rewarded, S0 arm choices unrewarded. Precisely, one reason we included this session-leading "anchor" of R1 confirmations in our experimental design was to enable a certain level of separation between behavioral response to *environmental* novelty and behavioral response to *rule* novelty. In EdRR 2, animals still tended to explore S0 surfaces early, but markedly less so than in EdRR 1 (figure 3c, bottom). 18/40 R1 trained animals (45%) chose the S0 arm on the first trial, 10/40 (25%) on the second

trial, 2/40 (5%) on the third trial, leaving 10/40 (25%) animals who did not explore an S0 arm in any of the initial three EdRR 2 trials, compared to only 1/40 (2.5%) in EdRR 1. The initial novelty of the EdRR environment thus appeared to increase exploratory behavior in two senses: firstly, "topologically" with respect to physical exploration of all accessible regions of the environment; secondly, "nomologically" (from Greek *nomos*, meaning law, rule, or custom) with respect to pointedly exploring beyond the bounds dictated by the R1 rule, an exploration which translates as an initial active preference for choosing S0 arms.

Explore first, whatever the expense: Finally, we looked at the conditional likelihoods, at the population level, of choosing the S1 or S0 arm in every trial of the first two EdRR sessions (with specific focus on the initial three trials; figure 3d, middle & bottom), comparing them to the S1 vs S0 population choices made during their final R1 training sessions (figure 3d, top). Here we saw that the likelihood of an animal from the R1 trained population choosing an S1 arm on the first EdRR 1 trial was 0.425, and 0.368 for an animal from the control group (figure 3d, middle). We tested the statistical significance of these results against both the 0.5 chance value expected if a population were choosing randomly and also against both populations' final R1 likelihoods of choosing the S1 arm on the first trial (Final R1, P(S1|1st trial); R1 trained = 0.825; controls = 0.474). These analyses revealed no significant difference compared with chance level in either population (t-tests with Welch correction against 0.5; R1 trained, $t(39) = 0.95$, $p = 0.349$; controls, $t(18) = 1.16$, $p = 0.262$) but a significant difference in the R1 trained population only when compared with their first trial behavior in the final R1 session (T-tests with Welch correction; R1 trained, $t(39) = 5.05$, $p < 0.0001$; controls, $t(18) = 0.93$, $p = 0.365$). On the second trial, however, the likelihood of choosing the S1 arm decreased further to 0.175 in the R1 trained population, and to 0.263 in the control group. As this observation indicates, out of those R1 trained animals who had chosen to explore the S0 arm on the first trial (i.e. 23/40, who were therefore not rewarded on this trial and thereby experienced a confirmation of R1 "in the negative", i.e. S0 + no-reward), a significant majority (18/23, 78.3%) nevertheless did not consequently choose the S1 arm on the second trial, but instead chose again to explore an S0 arm, foregoing the possibility of reward even under the modalities of the R1 rule. In control animals also, 10/12 (83%) who chose S0 on the first EdRR 1 trial (and were thereby unrewarded) chose S0 again on the second trial. Out of those R1 trained animals who chose the S1

arm on the first trial, thereby receiving a reward and a "positive" confirmation of R1 (i.e. S1 + reward), again the most significant proportion (15/17, 88.2%) nevertheless chose to explore the S0 arm on the second trial. In control animals, 4/7 (57%) who had chosen S1 on the first trial chose S0 on the second trial, commensurate with both random choice behavior and control group choice behavior in the final R1 session. This low trial 2 likelihood of S1 choice was significantly different from chance level in both the R1 trained population ($t(39) = 5.34$, $p < 0.0001$) and in controls ($t(18) = 2.28$, $p = 0.035$), but significantly different from the S1 choice likelihood on the same trial in the final R1 session only in the R1 trained population ($t(39) = 8.6$, $p < 0.0001$; controls, $t(18) = 2.03$, $p = 0.057$). It was only on EdRR 1 trial 3 that overall population level decision behavior shifted in the R1 direction, with a likelihood in R1 trained animals of 0.75 and in control animals of 0.734 of choosing the S1 arm. Even at the extremes of individual behavior, only 1 R1 trained animal chose to explore S1 arms on each of the initial presentations of the three pairs in EdRR 1, compared to 3 who chose three consecutive S0 arms. This serves to underline the cognitive implication of all the above results: under the initial effect of high novelty, the drive to explore overcame even at the cost of repeated foregone reward.

During the rest of EdRR 1, the R1 trained population level likelihood of choosing S1 stayed above 0.5 on every trial except the final one, a trial of the simplest level of EdM complexity (level 0, see below), whereas in the control group this surface choice likelihood oscillated more or less evenly above and below the 0.5 chance level (see below). Compatible with our hypothesis of decreasing impact of novelty with repeated exposure to the novel environment, in EdRR 2 we did not observe the same low S1 choice values in the initial trials (figure 3d, bottom). The probability of an animal from the R1 trained population choosing the S1 arm in either of the first two trials was higher (0.575 in both) than in EdRR 1 and not significantly different to chance likelihood level in either case. Similar to EdRR 1, however, the likelihood of the R1 trained population choosing an S1 arm subsequently stayed above 0.5 throughout the remaining trials of EdRR 2. Indeed, in both EdRR 1 and EdRR 2, overall surface choice behavior was significantly different to chance level in the R1 trained population (T-test with Welch correction and unbiased Cohen effect sizes; EdRR 1, $t(22) = 5.05$, $p < 0.0001$, $d = 1.05$; EdRR 2, $t(22) = 7.42$, $p < 0.0001$, $d = 1.54$; note that the effect size of this difference is ~1.5 times greater in EdRR 2 compared to EdRR 1) but not in controls (EdRR 1, $t(22)$

= 0.68, p = 0.5; EdRR 2, $t(22)$ = 1.96, p = 0.06). It is also clear from figure 3d (bottom) that troughs in the R1 trained population likelihood of choosing S1 preferentially occurred on trials of complexity level 0, which leads us onto our next focus of analysis.

3. More complex EdM trials occasion higher R1 interference in choice behavior.

Based on previous work from our lab (Al Abed et al., 2016), we expected performances in the EdRR task to be a function of trial complexity, with scores highest on the least complex trials (level 0). To recall (see Materials & Methods), trials at complexity level 0 are most closely equivalent to basic spontaneous spatial alternation tasks, since the pair presented during such trials is the same pair that was presented in the directly previous trial, i.e. without any interposed presentations of other pairs. Nevertheless, it must also be recalled that, *unlike* with the basic T- or Y-maze apparatus, even on spontaneous alternation-like EdM trials of complexity level 0, animals are holding other context-relevant cognitive content relating to the two other arm-pairs, uncertain of when they will next need it.

EdM performance as a function of trial complexity: In order to achieve stronger statistical power in our trial complexity level analyses, session performances were aggregated into blocks of three, according to our experimental design whereby each block was conceived so as to comprise exactly 12 trials of each level of complexity (for 5 levels of trial complexity, giving a total of 60 trials, excluding initial presentation of each pair in each session). At this level of analysis, we observed that, as with naïve mice from previous studies (Al Abed et al., 2016), the percentage of incorrect EdM responses in both the R1 trained and control populations was a function of trial complexity (figure 4a; top, three-way ANOVA, major effect of 'Difficulty', $F(4, 3500) = 130$, $p < 0.0001$). At this level also, it was clear that only in block 1 (figure 4a, top left) and only on the most complex trials, did the R1 trained population make significantly more errors than controls (pairwise t-tests with Bonferroni correction, between 'Group' difference; complexity level 3, $t(103)$ = 3.01, p = 0.016; complexity level 4, $t(105)$ = 3.02, p = 0.016).

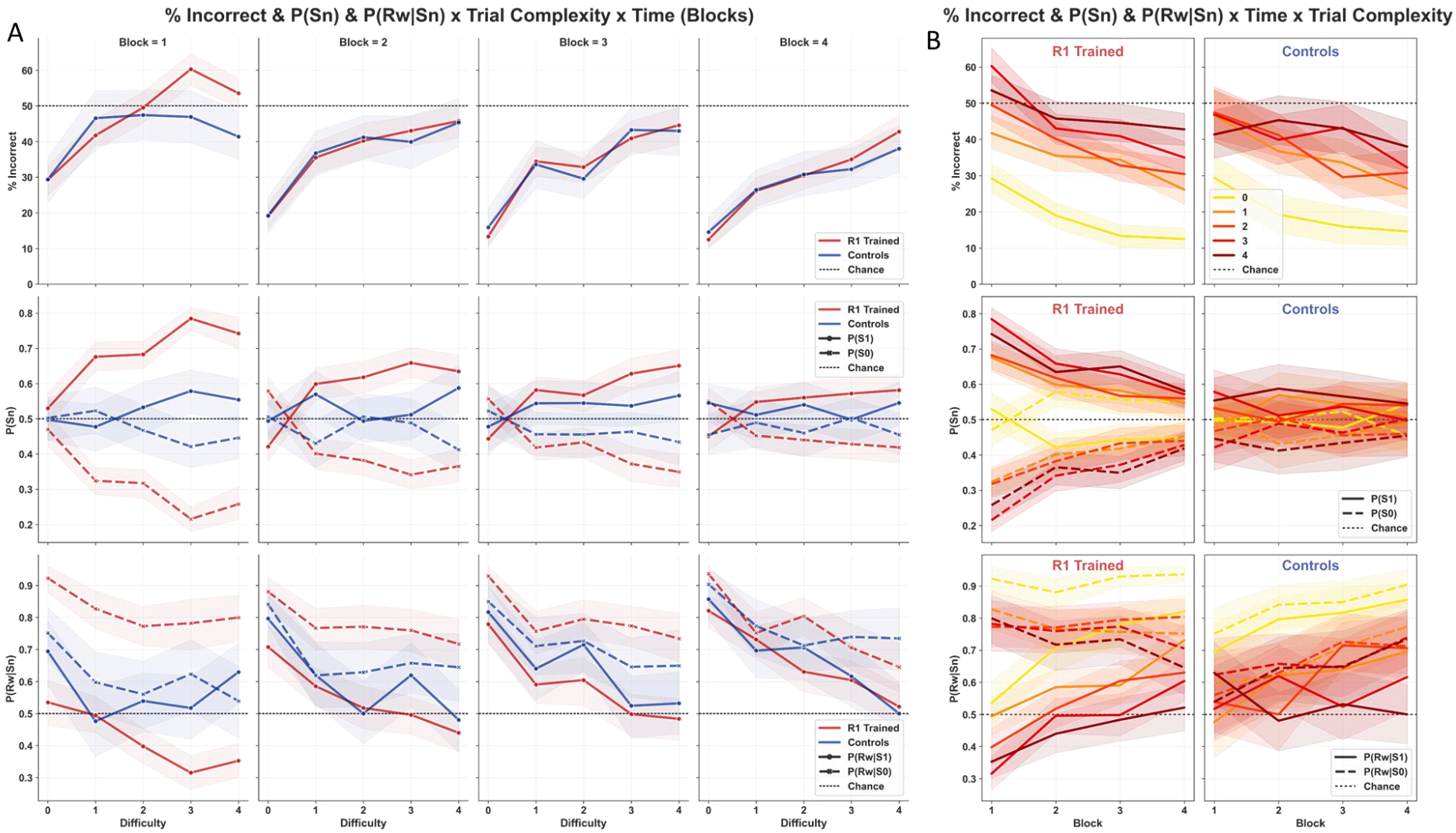

A
% Incorrect & P(Sn) & P(Rw|Sn) x Trial Complexity x Time (Blocks)
Block = 1
Block = 2
Block = 3
Block = 4
% Incorrect
P(Sn)
P(Rw|Sn)
Difficulty
R1 Trained
Controls
P(S1)
P(S0)
P(Rw|S1)
P(Rw|S0)
Chance
B
% Incorrect & P(Sn) & P(Rw|Sn) x Time x Trial Complexity
R1 Trained
Controls
0
1
2
3
4
Block

Figure 9 - R1 interference in EdRR very significantly impacted by trial complexity.
By aggregating sessions into blocks of 3, each block contains exactly 12 trials of each trial complexity level, allowing for more robust statistical analysis while still maintaining a sufficient dimension of evolution over time. Elsewhere, we analyze all trials, averaged across complexity levels and including initial trials (see figure 5). Here we analyzed errors, probability of choosing S1 or S0 ('P(Sn)') and probability of being rewarded (outcome) given that S1 or S0 had been chosen ('P(Rw|Sn)'), all as a function of trial complexity. We provide two visualizations of the same findings. In (A), trial complexity from 0 to 4 is represented on the x-axis and each column represents a block of EdRR, from 1 to 4. In (B), EdRR blocks are represented on the x-axis, values according to trial complexity are represented as color-coded curves, from yellow (level 0) to burgundy (level 4), and each column represents one of the populations: R1 trained, left; controls, right. Error bands in all cases represent 95% confidence intervals. (**A**) + (**B**), top row: % of EdM errors was significantly impacted by trial complexity, matching what we expected based on the existing classical EdM task literature. Level 0 performance in both groups was particularly strong, even in block 1. By contrast, R1 trained animals performed significantly worse than controls only on level 3 + 4 trials and only in block 1. Their performance in these trials was below what could be explained by chance. In subsequent blocks, EdM performance in both groups improved as a function of time, nevertheless always as a function of trial complexity. (**A**) + (**B**), middle row: P(Sn) expressed as a probability between 0 and 1. P(S1) increased highly significantly as a function of trial complexity in R1 trained animals. This dependence of P(Sn) on trial complexity decreased gradually but was still visible and significant in all blocks. P(Sn) on level 0 trials thus demonstrated the least impact of R1 training. Indeed, a reflection of high EdM performance on level 0 trials is precisely that R1 trained animals were capable of actually choosing S0 more often than S1, but on these easiest trials only (more details in text). In controls, by contrast, P(Sn) fluctuated on all trial complexity levels except level 4 where a slightly but consistently higher P(S1) value was observed, potentially indicating a consistent response strategy involving surface on these trials. (**A**) + (**B**), bottom row: P(Rw|Sn) expressed as a probability between 0 and 1. The probability of R1 trained animals being rewarded was significantly higher on trials of all complexity levels when they chose S0 as opposed to S1. This is a reflection of the fact that the majority of their EdM errors were committed on S1 arms, implying that reward was most often "waiting" for them on the S0 arm of a pair. P(Rw|S0) did not significantly evolve over time, except on level 4 trials, where it decreased slightly. Rather, with repeated EdRR training, P(Rw|S1) gradually rose up to almost comparable levels as P(Rw|S0) on all complexity levels, though even this improvement with time had a trial complexity dependence most easily visualized in block 3 where P(Rw|S1) decreases step like from 0.8 at level 0, to 0.6 on levels 1-2, to 0.5 on levels 3-4. Whereas, in contrast, P(Rw|S0) is higher on level 0, as expected, but almost equal at ~0.75 on all trial complexity levels.

Errors made by the R1 trained population on complexity level 3 trials in block 1 went significantly beyond chance level, i.e. the percentage of errors that would be achieved by choosing randomly (i.e. 50%), and a similar but not significant trend was seen with errors made on complexity level 4 trials (independent t-tests; difficulty level 3, $t(119) = 4.29$, $p < 0.0001$; difficulty level 4, $t(119) = 1.61$, $p = 0.1$). We also observed that the number of errors decreased significantly and consistently across all complexity levels as a function of time/repeated training (three-way ANOVA, major effect of 'Block', $F(3, 3500) = 72$, $p < 0.0001$). In figure 4b (top, left and right) we have given an alternative graphical representation of these same evolutions. Here, time in blocks of training is represented on the x-axis, performance curves for each level of trial complexity are color-coded on a heat-scale from yellow (level 0) to burgundy (level 4), and the two populations are assigned to separate columns; R1 trained population on the left, controls on the right. This representation renders even clearer just how significantly fewer errors occurred on complexity level 0 trials (yellow lines), those corresponding closest to spontaneous alternation tasks, across all training blocks and in both the R1 trained and control populations (pairwise t-test between % errors performed at trial complexity level 0 and level 1, averaged across groups and blocks, $t(58) = 5.47$, $p < 0.0001$).

Surface choice probability as a function of trial complexity: In figure 2c above, we measured R1 interference on EdM performance as a function of bias observable in the nature of the errors committed during performance of the EdRR task. This demonstrated that although R1 trained animals and control animals made comparable overall numbers of errors, the nature of these errors was significantly and persistently biased towards S1 in R1 trained animals. However, such an analysis presented alone may give rise to the misleading idea that bias exists only when mistakes are made, whereas, in fact, bias induced by previous learning cannot and should not be reduced to something that manifests only when errors arise, since errors are as much a function of the environment as they are of the cognitive processes of the organism. Crucially, in line with the origin of the term in the game of bowls, where there is a bias in cognition, we should expect it to manifest independently of whether a given subject-environment interaction gives rise to a correct or to an incorrect outcome (see Discussion below). For this reason, we also measured the overall trial-by-trial probability of choosing S1 or S0, independently of whether or not that choice led to an EdM error (figure 4a & figure 4b, middle rows).

In a behavioral paradigm based on spatial alternation, in which the space in question is divided half and half by two different surfaces, all other things being equal besides we should expect the probability of exploring each surface type to converge towards 0.5, with a certain calculable margin of variance, in both animals performing the task perfectly and in animals performing purely randomly. Comparing this to our results from each block (figure 4a, middle row), what we observed in controls is that the mean choice probability of exploring each surface ('P(S1)' and 'P(S0)') lay between maximal extremes of 0.6 and 0.4, respectively, across all blocks and across all levels of trial complexity, with the 95% confidence interval (CI) error bands over-lapping between the two surfaces to the extent that the mean probability of choosing one surface over another achieved statistical significance at only three discrete points; level 3 complexity trials in block 1 (pairwise t-test with Bonferroni correction, $t(112) = 3.42$, $p = 0.004$), and levels 1 and 4 complexity trials in block 2 (pairwise t-tests with Bonferroni correction, $t(112) = 2.91$, $p = 0.02$; $t(112) = 3.61$, $p = 0.002$, respectively). This result indicated a dependent interaction between trial complexity and surface choice in the control group, which emerged as statistically significant on complexity levels 3 and 4 only when averaged across all blocks (pairwise t-tests with Bonferroni correction, interaction 'Difficulty*P(Sn)', level 3, $t(454) = 2.74$, $p = 0.031$; level 4, $t(454) = 5.33$, $p < 0.0001$, respectively). These same observations are also visible in figure 4b, middle row, right column; moving from block to block along the x-axis, we see that the bounds of the vertical 95% CI variance in surface choice probability values did not go significantly beyond 0.6 and 0.4. We did, however, again see the same visible but limited tendency for even control animals to choose S1 more than S0 in trials above complexity level 0, most prominently so in complexity level 4 trials. In a two-way ANOVA averaged across all blocks, this gave rise to a significant global effect of surface on choice in controls, with a significant interaction between difficulty and surface ($F(1, 2270) = 34$, $p = 0.001$, with Tukey HSD post-hoc; $F(4, 2270) = 3.42$, $p = 0.009$, respectively).

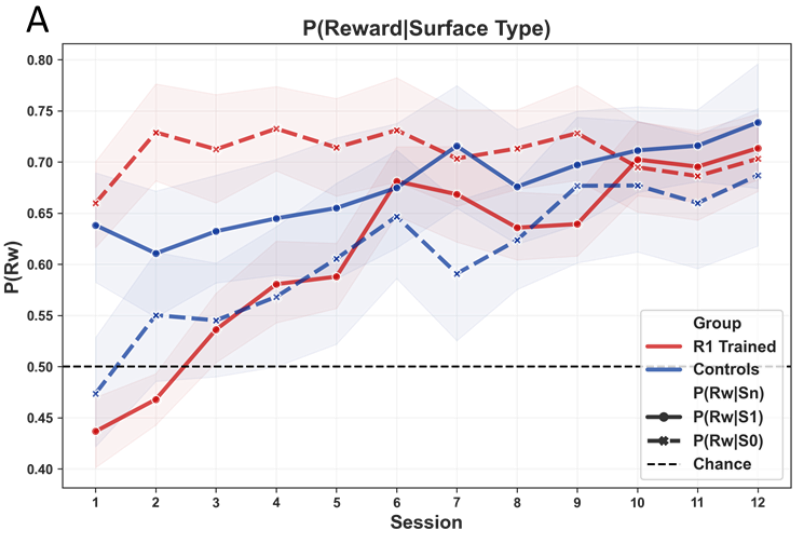
Although we observed a surface effect in controls which could not be explained by prior learning (analyzed further below), the R1 trained population was significantly more impacted by surface (two-way ANOVA, interaction 'Group*P(Sn)'; $F(1, 7076) = 104$, $p < 0.0001$). In this population, we observed a very strong dependent interaction between trial complexity and surface choice. Across all blocks, the probability of choosing S1 increased highly significantly (mirrored by a proportional decrease in S0 choice

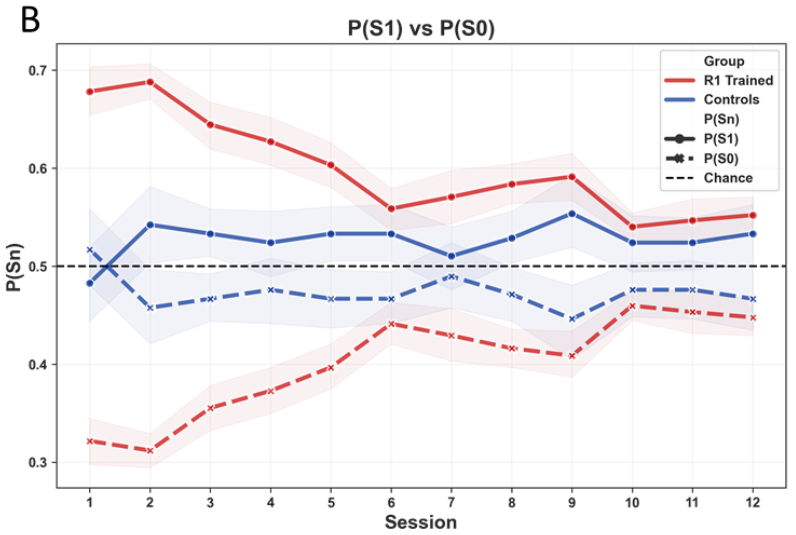
probability) as a function of trial complexity (pairwise t-tests with Bonferroni correction, $t(958) > 4$, $p < 0.0001$ for all levels; unbiased Cohen effect sizes, level 0, d = 0.3; level 1, d = 0.85; level 2, d = 0.86; level 3, d = 1.31; level 4, d = 1.15), though the magnitude of the difference did decrease gradually as a function of time/repeated training (two-way ANOVA, interaction 'P(Sn)*Block'; $F(3, 4792) = 69$, $p < 0.0001$; ANOVA with Tukey post-hoc on P(Sn) in block 4; $F(1, 1198) = 35.2$, $p = 0.001$; unbiased Cohen effect size, d = 0.34). In fact, in the visual comparison seen in supplementary figure S.2e, we can see that P(S1) on complexity level 3-4 trials (designated 'complex' in this graphical representation) during the first block of EdRR were almost as high as P(S1) in the final R1 session, despite the fact, just seen, that the majority of these P(S1) choices in EdRR block 1 were giving rise to unrewarded errors.

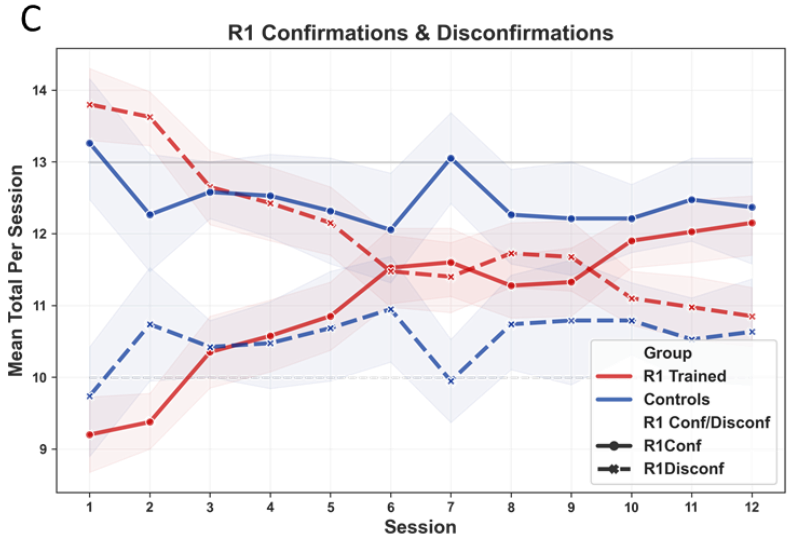
P(S1) and P(S0) values were most closely matched in the R1 trained population on complexity level 0 trials in block 1 (mean values P(S1) = 0.53, P(S0) = 0.47). In subsequent blocks, the R1 trained population was significantly more likely to choose S0 arms in complexity level 0 trials. Although this S0 preference at level 0 may seem counter-intuitive, it is actually easily explained by the fact that the same animals were significantly more likely to choose S1 arms at all other trial complexity levels, meaning that whenever a complexity level 0 trial occurred, there was a statistically higher probability that the animal had chosen the S1 arm on the immediately previous presentation of the same arm pair. Thus, if the animal spatially alternated on a level 0 trial (which, referring to figure 4a, top, we can see that the R1 trained population was doing on average 80% or more of the time in these blocks), this would necessarily imply choosing the S0 arm significantly more often than the S1 arm, only on level 0 trials. Still, that R1 trained animals had the cognitive capacity to choose S0 more often on these level 0 trials is note-worthy.

In figure 5b, we have represented session-by-session P(S1) vs P(S0) values averaged across all trial complexity levels, this time also including the initial trials, which have no EdM trial complexity level. And, in supplementary figure S.2a, for completeness, we have also isolated and represented the session-by-session P(S1) vs P(S0) values for these initial trials only, where we can again see the dramatic change in P(Sn) between the high novelty EdRR 1 and subsequent lower novelty EdRR sessions in R1 trained animals, as discussed above (figure 4a-d).

Reward probability, given surface choice, as a function of trial complexity: Next, we looked at the resultant conditional probabilities for animals from both groups to obtain a food reward (correct EdM response) given the surface chosen ('P(Rw|Sn)') and as a function of trial complexity level (figure 4a and figure 4b, bottom rows). Since we are looking at the probability of obtaining a reward as a function of EdM trial complexity level, the initial presentation of each pair in each session is not included, as per the experimental design of the classical EdM task (see Materials & Methods). Thus, we recall again that in these initial pair presentations P(Rw|S1) = 1 and P(Rw|S0) = 0, for both the R1 trained and control populations (these will be accounted for in figure 5a below). In contrast to P(Sn), where both perfect and purely random EdM performances would result in the values of P(S1) and P(S0) converging towards 0.5, a hypothetical perfect EdM performance in the EdRR environment would result in P(Rw|S1) = 1 and P(Rw|S0) = 1, while a purely random performance would result in the values of P(Rw|S1) and P(Rw|S0) converging towards 0.5. A general trend for the probability of reward on both surfaces to increase across time/repeated training was observed, with no significant difference between the two groups in overall P(Rw) (one-way ANOVA, $F(1, 6473) = 0.38$, $p = 0.54$; this is a logical extension of overall EdM performances being similar between the two groups, as seen in figure 3a above). There was, however, a highly significant interaction between 'Group' and 'P(Rw|Sn)' (two-way ANOVA with Tukey HSD post-hoc; 'Group*P(Rw|Sn)', $F(1, 6471) = 82$, $p = 0.001$). Analyzing within the groups, P(Rw|S1) was very significantly lower than P(Rw|S0) in the R1 trained population (figure 4a, bottom; one-way ANOVA with pairwise Tukey HSD post-hoc, $F(1, 4340) = 529$, $p = 0.001$, Cohen effect size, $d = 0.699$), an expected result given that their P(S1) was significantly higher than their P(S0). This significant difference between P(Rw|S0) and P(Rw|S1) values in the R1 trained population was observed at all trial complexity levels across blocks 1 – 3 and persisted, albeit less significantly, in block 4 also (pairwise t-tests with Bonferroni correction averaged across all blocks; 'P(Rw|Sn)*Difficulty', $p < 0.0001$ at all difficulty levels).







Table

	EdRR Correct (rewarded)	EdRR Incorrect (unrewarded)
S1	R1 confirmation (+)	R1 disconfirmation (-)
S0	R1 disconfirmation (+)	R1 confirmation (-)

Figure 10 - Controls but not R1 trained animals update P(Sn) in accordance with P(Reward|Surface Type).

Overall per session P(Rw|Sn) values, averaged across all complexity levels and encompassing initial per session trials on each pair, in which P(Rw|S1) = 1 and P(Rw|S0) = 0. P(Sn) per session is represented beside this, allowing easy visualization of the contrast between the R1 trained and control groups with respect to the relationship between, specifically, P(Rw|S1) and P(S1). All error bands represent 95% confidence intervals (for detailed statistical analyses, see main text). (**A**) P(Rw|Sn). The R1 trained and control populations experienced opposing outcome values in this regard. In R1 trained animals, P(Rw|S0) was higher than P(Rw|S1) throughout the first 3 blocks. Their P(Rw|S0) did not significantly evolve over repeated training, rather P(Rw|S1) gradually increased over time until the two values came level in the final block only. By contrast, in controls, P(Rw|S1) was slightly but robustly higher than P(Rw|S0) across all 12 sessions of EdRR, a contribution to which comes from the initial per session trials in which P(Rw|S1) = 1. (**B**) P(Sn). The fact that the control group displayed a robust trend, beginning only from EdRR 2, to slightly favor S1 thus seems to find at least partial explanation in this group's robustly higher P(Rw|S1) values. It is also clear to what extent P(S1) values in the R1 trained population remained persistently resistant to the fact that they were being rewarded less often when they chose S1, and precisely more often when they chose S0. (**C**) + (**Table**) Mean total R1 confirmations and disconfirmations per session per group. Our EdRR protocol was designed specifically in such a way that trial-by-trial outcomes could be qualified as either confirmations or disconfirmations of the previously learned R1 rule. In the Table, we recap briefly which EdRR outcomes constitute an R1 confirmation and which a disconfirmation. Recalling that in EdRR the outcome of the initial per session trial on each pair is always a R1 confirmation, this implies that both hypothetical perfect and perfectly random EdRR performances will generate a total number of R1 confirmation outcome trials converging towards 13 (out of 23) and of R1 disconfirmation outcome trials converging towards 10. The horizontal unbroken and dashed light grey lines at 13 and 10, respectively, are visual anchors to help represent this. Thus, it is visually clear that total numbers of confirmations and disconfirmations in control animals did tend to converge towards these modelled values, whereas in the R1 trained population initial values were inverted in this respect; many more R1 disconfirmations being the result of many incorrect (in terms of EdM) S1 choices. Only during the final block did this population begin to converge towards modelled values. Referring back to (B), the visible lack of decrease in P(S1) between EdRR 1 and EdRR 2 in the R1 trained population is all the more striking in light of the number of R1 disconfirmations we see they had experienced during EdRR 1.

In controls also, a trend for P(Rw|S0) to be higher than P(Rw|S1) emerged starting in block 2, but reached statistical significance only in block 4 in complexity level 4 trials (pairwise t-tests with Bonferroni correction averaged across all blocks, interaction 'P(Rw|Sn)*Difficulty'; levels 0-2, $p > 0.9$; level 3, $p = 0.34$; level 4, $t(106) = 3.3$, $p = 0.007$). Importantly, however, we will now see that, in controls, overall surface-based differences in P(Rw|Sn) and P(Sn) were commensurate with each other, which was not the case in the R1 trained population (figure 5a-b).

It was essential to also analyze overall P(Rw|Sn), including the initial three trials from each session on which P(Rw|S1) = 1 and P(Rw|S0) = 0, since the outcome of these trials would exert a strong influence on each animal's summed P(Rw|Sn) experience per session. Indeed, what we observed (figure 5a) is that when these initial trial P(Rw|Sn) values were included, P(Rw|S1) and P(Rw|S0) were drawn closer together in the R1 trained population than when they were analyzed according to EdM trial complexity, coming into almost perfect alignment by the final three sessions/final block. In controls, interestingly, inclusion of the initial trials revealed that their P(Rw|S1) was globally higher than P(Rw|S0) in all EdRR sessions (figure 5a), a fact which was hidden when these values were analyzed according to EdM trial complexity only. Hence, looking at figure 5a and 5b as we have presented them here side-by-side, we are invited to entertain the possibility that the observed trend for P(S1) values to rise slightly higher than P(S0) values as a function of time in controls may at least in part be explained as a cognitive behavioral response (e.g. some kind of putatively Bayesian updating) to this population's consistently higher P(Rw|S1) values. Lending weight to this hypothesis of an ongoing P(Rw|Sn) updating in controls is the observation that P(Rw|S1) was already higher than P(Rw|S0) in EdRR 1, whereas P(S1) rose above P(S0) only starting from EdRR 2 (figure 5a-b, see Discussion).

4. Hallmarks of myside confirmation bias during EdRR.

The specific objective behind the design of our behavioral paradigm was to model state-action policy revision in mice in a real-world, everyday-like environment, i.e. where the situations in which we have to employ acquired cognitive content may vary greatly in complexity and uncertainty, and where new information coming from the environment is often ambiguous with respect to previously acquired cognitive content, such as beliefs,

rules, and state-action policies. With that in mind, we designed our paradigm in such a way that performance of the EdRR task would generate, trial-by-trial, a sequence comprised of both confirmations and disconfirmations of the previously learned R1 tactile discrimination rule (see Materials & Methods). In these EdRR task conditions, R1 *confirmations* could take two forms; correct (i.e. rewarded) EdM response on an S1 arm, or incorrect (i.e. unrewarded) EdM response on an S0 arm. Likewise, R1 *disconfirmations* could also take two forms; correct EdM response on an S0 arm, or incorrect EdM response on an S1 arm (see figure 5, table). Recall also that on initial per session presentation of each pair, arm choices can be neither correct nor incorrect in terms of the classical EdM task (which is based on making a choice on trial N according to the choice made on the same pair on trial N-1, thus excluding the initial N which has no corresponding N-1). In EdRR, upon initial per session presentation of each pair, only S1 arms were rewarded, meaning that regardless of whether S1 or S0 was chosen, response on these initial three trials always gave rise to an R1 confirmation. In other words, the optimal action policy for the EdRR task can be formalized as follows: "1. On the initial trial of each pair in a given session, choose the S1 arm. 2. On all subsequent pair presentations, spatially alternate with respect to previous choice." One of the questions we wanted to investigate with this experimental design was whether, during the cognitive process of rule revision, we would observe behavior that reflected a specific myside confirmation bias relative to the previously internalized R1 rule. Indeed, a central hypothesis of our experimental design was that precisely such a bias would be observed in these conditions.

Session-by-session R1 confirmations and disconfirmations: In figure 5a-b, we see evidence that, in the R1 trained population but not in controls, there was a lag in rule revision observable as a delay in updating P(Sn) as a function of P(Rw|Sn). Indeed, in figure 5, we have presented the all-trial inclusive, session-by-session, trial complexity independent representations of these P(Sn) and P(Rw|Sn) values side-by-side in order to make this lag clear. To these, we have added a representation of the session-by-session mean total number of R1 confirmations and disconfirmations experienced during the second phase EdRR task by animals from each of the two groups (figure 5c). As a heuristic, we have also included a table that recapitulates which EdRR trial-context-specific outcomes constitute a confirmation and which a disconfirmation of R1. During the first five EdRR sessions, we observed that the R1 trained and control populations

had diametrically opposed experiences in terms of R1 confirmations and disconfirmations; the R1 trained population experienced significantly more disconfirmations than confirmations, while the controls experienced significantly more confirmations than disconfirmations. We additionally added two extra elements to figure 5c in order to better frame these results: since the initial per session trial on each pair necessarily constitutes an R1 confirmation, this entails that, during EdRR, both a perfect EdM performance and a purely random EdM performance will tend to elicit a per session number of R1 confirmations converging towards 13 (solid light grey horizontal line at y = 13) while disconfirmations will correspondingly converge towards 10 (dashed light grey horizontal line at y = 10). Visually, these elements allow us to see that, in controls, R1 confirmation and disconfirmation averages remained close to these modelled values throughout EdRR, whereas in the R1 trained population, in the early sessions, the average total R1 confirmation and disconfirmation values were inverted with respect to the same modelled values (mean values EdRR 1, R1 confirmations = 9.2; R1 disconfirmations = 13.8). Across the 12 EdRR sessions, these values came close to the modelled values in the R1 trained population only during the final block, but even then still not as close as in controls. Thus, while both the R1 trained and control populations' R1 confirmation and disconfirmation values were significantly different to the modelled values of 13 and 10, respectively, this difference was significantly greater in the R1 trained population (T-tests with unbiased Cohen effect sizes; R1 trained, $t(479) = 24.6$, $p < 0.0001$, $d = 1.12$; controls, $t(227) = 5.2$, $p < 0.0001$, $d = 0.34$; one-way ANOVA with Tukey HSD post-hoc between 'Group', $F(1, 706) = 112.5$, $p = 0.001$). Finally, it is of particular interest, in terms of rule revision, to note that there was no decrease in P(S1) between sessions 1 and 2 in R1 trained animals (figure 5b, pairwise t-test, $t(39) = 0.56$, $p = 0.5$). This relative absence of revision of P(S1) in EdRR 2 is to be considered in light of the average 13.8 R1 disconfirmations experienced during EdRR 1, accounting for 60% of all trials.

5. R1 bias persists in cognitive behaviors over extensive training in the EdRR environment.

Population differences in decision latency fade during EdRR: Prior to their introduction into the EdRR environment, in the latter half of the R1 training phase, the R1 trained

population displayed significantly higher overall decision latency values compared to controls, and also significantly higher within group decision latency values on S1 arms compared to S0 arms (Stevens et al., 2022a). Upon introduction to the EdRR environment, we observed an initial increase in overall decision latency in both the R1 trained and control populations compared to their final R1 session values (figure 6a; R1 trained, ~2900 ms up from ~2500 ms in final R1; controls, ~2600 ms up from ~2000 ms in final R1). However, overall decision latency remained significantly higher in the R1 trained population only during EdRR 1, after which there was no longer any significant difference in overall decision latency between the two populations (decision latency comparison between groups; pairwise t-tests with Bonferroni correction and unbiased Cohen effect sizes; EdRR 1, $t(115) = 3.38$, $p = 0.01$, $d = 0.54$; EdRR 2, $t(92) = 2$, $p = 0.64$). We did observe that mean decision latency was higher in the R1 trained population during EdRR 2 when choosing S0 arms compared to S1 arms, having actually increased from EdRR 1 to EdRR 2, such that averaged across the three EdRR sessions of block 1, the R1 trained population had higher decision latencies when their decision-making led them to explore S0 arms, which is to say, specifically when they were deciding to "transgress" the R1 rule (figure 6a, difference approached statistical significance; pairwise t-tests with Bonferroni correction and unbiased Cohen effect sizes, $t(238) = 2.5$, $p = 0.055$, $d = 0.3$). From block 2 onwards, however, there were no further observable differences in decision latency, neither by surface chosen nor, in important contrast to the R1 phase, by experimental group. These results can be interpreted as reflecting initially increased cognitive conflict in overcoming R1 in the EdRR environment (increased decision latency on S0 choices compared to S1). Furthermore, the fact that decision latencies in both groups decreased steadily over repeated EdRR training also reveals that, overall, the spatial alternation "shift" behavior elicited by the EdM task does indeed come more spontaneously to mice than the "win-stay" S1 exploitation behavior demanded by the previous R1 rule (Stevens et al., 2022a). That this should be the case even when the EdM task is being conducted under rule revision conditions producing significant effects in many other behaviors is all the more striking.

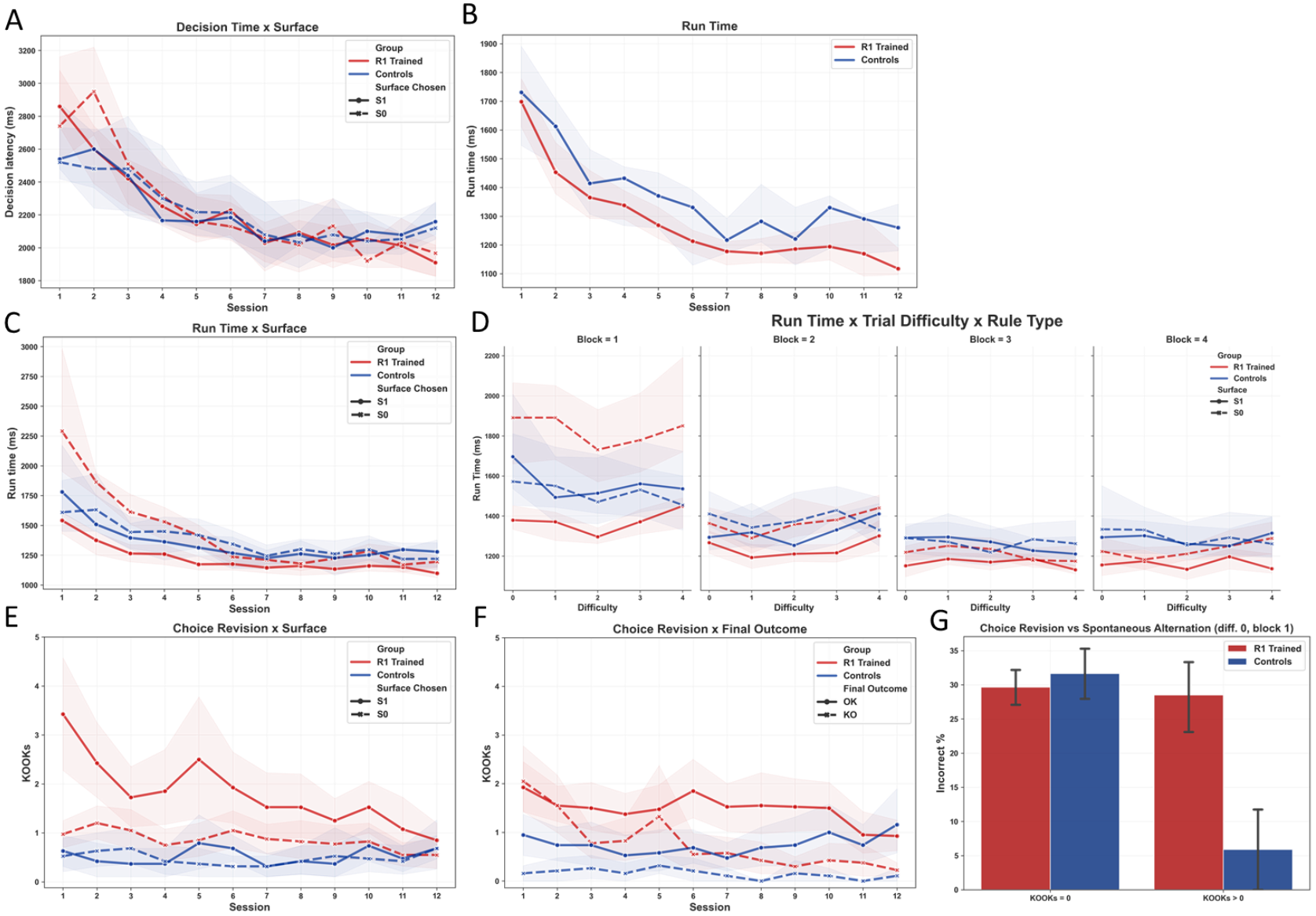

A
Decision Time x Surface
Decision latency (ms)
Session
Group
R1 Trained
Controls
Surface Chosen
S1
S0
B
Run Time
Run time (ms)
Session
R1 Trained
Controls
C
Run Time x Surface
Run time (ms)
Session
Group
R1 Trained
Controls
Surface Chosen
S1
S0
D
Run Time x Trial Difficulty x Rule Type
Block = 1
Block = 2
Block = 3
Block = 4
Run Time (ms)
Difficulty
Group
R1 Trained
Controls
Surface
S1
S0
E
Choice Revision x Surface
KOOKs
Session
Group
R1 Trained
Controls
Surface Chosen
S1
S0
F
Choice Revision x Final Outcome
KOOKs
Session
Group
R1 Trained
Controls
Final Outcome
OK
KO
G
Choice Revision vs Spontaneous Alternation (diff. 0, block 1)
Incorrect %
KOOKs = 0
KOOKs > 0
R1 Trained
Controls

Figure 11 - Confirmation bias-like effects highly persistent in post-decision cognitive behaviors.

To allow deeper investigation of deliberative processes, not only before but also after trial choice, we developed analyses for measuring decision latency, post-decision run time, and choice revision behaviors, thus making the most of behaviors that are uniquely possible with the radial maze apparatus. These behaviors proved to be highly revelatory with respect to R1 interference during EdRR and offered powerful insights into the cognitive and putative neurobiological mechanisms underpinning the cognition of rule revision. All error bands represent 95% confidence intervals (see main text for detailed statistical analyses). (**A**) Median decision latencies (in milliseconds) by surface and by session. Both populations displayed a significant time-dependent decrease in decision latency as a function of repeated EdRR training. In EdRR 1 + 2, R1 trained animals displayed a trend for significantly higher decision latency, particularly on S0 choice trials in EdRR 2, putatively an indication of cognitive effort to inhibit the R1 drive to choose S1. (**B**) Median run times (in milliseconds), independent of surface, by session. Averaged across all EdRR sessions, R1 trained animals displayed lower post-choice run times, an indicator of post-choice confidence, than controls. This was in contrast to the R1 training phase during which controls displayed lower overall post-choice run times. (**C**) Median run times (in milliseconds) by surface chosen. R1 trained animals displayed significantly higher per session S0 run times up until EdRR 5 (note the lag in group homogenization compared with evolution of decision latency). Furthermore, R1 trained run times on S1 arms remained consistently lower, indicating a certain level of persistent "over-confidence" in these choices, up until session 12. (**D**) Median run times (in milliseconds) by difficulty by block. Confirming the robustness of these lower S1 run times, we can see here that it was reliably the case over all trial complexity levels and in all blocks. (**E**) Mean total choice revision by surface by session. R1 trained animals displayed around twice as much choice revision behavior as controls, with a significant tendency for the final decision to be an S1 choice. This was especially true in early EdRR sessions, decreasing gradually as a function of time, but with a persistent trend still even in EdRR 12. (**F**) Mean total choice revision by outcome by session. Controls were significantly more likely to engage in rectifying rather than error-inducing choice revision across all EdRR sessions. Choice revision in R1 trained animals in EdRR 1 + 2 was as likely to lead to error as to a correct response. Only from EdRR 6 onwards was choice revision in R1 trained animals significantly more likely to be rectifying. (**G**) In level 0 trials, almost all choice revision in controls was rectifying whereas R1 trained animals were as likely to make an error with as without choice revision. This was driven by R1 bias (see main text).

Diminishing yet persistent differences in run time between populations and between surfaces: Next, we looked at post-choice run times, the trial-by-trial time taken to travel from the threshold between the central platform and the chosen arm to its distal, reward distributor containing zone. This post-choice fact of the radial maze apparatus is a central feature enabling analysis of cognitive processes in mice while they are physically realizing their choice, in a manner not possible where execution is quasi-instantaneous (e.g. lever press). Whereas in the latter half of R1 training, overall run time was higher in the R1 trained population (which we suggested may be due to control animals attaining higher confidence in being rewarded on every trial and therefore manifesting less post-choice hesitation (Stevens et al., 2022a)), in the EdRR task we observed that overall run time, averaged across all sessions and surfaces, was actually slightly but significantly lower in the R1 trained population (figure 6b, one-way ANOVA with Tukey HSD post-hoc, $F(1, 705) = 10.5$, $p = 0.001$). Looking at the constitution of these overall values, we observed that, during the first five EdRR sessions, run times were significantly higher in R1 trained animals on S0 arms compared to both control animals (figure 6c; see 95% CI error bands; this difference translated into a statistically significant overall between-group difference in S0 run times; one-way ANOVA with Tukey HSD post-hoc, $F(1, 705) = 7.6$, $p = 0.006$) and to their own S1 arm run times (figure 6c, one-way ANOVA with Tukey HSD post-hoc, $F(1, 956) = 77.6$, $p = 0.001$: no significant difference between S1 and S0 run times in controls; $F(1, 454) = 0.2$, $p = 0.67$). In fact, this within-R1 trained population, between-surface difference was significant even when we analyzed only blocks 3-4, i.e. even when there was no longer any significant session-by-session difference (one-way ANOVA with Tukey HSD post-hoc, $F(1, 478) = 6.6$, $p = 0.001$). Taking all this together, it seems that lower S1 run times in the R1 trained population throughout EdRR was driving this group's overall, surface-independent lower run time values. Interestingly, when we looked at these group level differences in run times as a function of trial difficulty and repeated training (figure 6d), we observed that, in the first block especially, run times were significantly greater in the R1 trained population on S0 arms at all trial complexity levels, including level 0, on which, as we have seen, these animals were nevertheless performing relatively well (figure 6d, pairwise t-tests 'Difficulty*Surface' with Bonferroni correction and unbiased Cohen effect sizes, within first block; level 0, $t(174) = 5.3$, $p < 0.001$, $d = 0.72$; level 1, $t(155) = 3.4$, $p < 0.001$, $d = 0.49$; level 2, $t(117) = 5.5$ $p < 0.001$, $d = 0.81$; level 3, $t(81)$

= 3.1, $p < 0.001$, $d = 0.55$; level 4, $t(82) = 4.2$, $p < 0.001$, $d = 0.76$). We can also see from figure 6d just how reliable, across trial difficulty and repeated training, the trend for lower S1 run times was in the R1 trained group. For perspective, we can compare this surface choice breakdown of run times to an EdM outcome (correct or incorrect) breakdown (supplementary figure S.3a, bottom). Here, we can see that with repeated training, run times on incorrect response trials increase more steadily at all trial complexity levels, but especially level 0, in the control group compared to the R1 trained group, i.e. towards a robust phenotype we also see in the classical EdM task for animals to have higher run times on incorrect response trials, putatively when they have a stronger feeling of reduced confidence in being rewarded (unpublished data, but see wildtype animals in supplementary figures S.6c and S.7d). A suggestion of these run time results is that, in R1 trained animals, there is some kind of highly persistent, putatively affective (McDonald et al., 2004; McDonald & Hong, 2004) over-confidence in being rewarded on S1 arms, even after extensive EdRR training throughout the majority of which P(Rw|S1) is in reality significantly lower than P(Rw|S0) (figure 5a).

Choice revision behavior: Finally, we also measured and analyzed physical choice revision behavior in both populations, i.e. when animals made an initial crossing of the threshold into one arm of a pair, but then revised their choice, returned to the central platform and made another choice. Once again, the possibility for this behavior is a feature of the radial maze apparatus. As a cognitive behavior, it is very similar to the deliberative rodent behavior referred to as "vicarious trial and error" (VTE) and thus putatively the result of representational contributions (Redish, 2016). However, since in the radial maze choice revision can also imply a certain level of advancing along an initially chosen arm, it may go beyond VTE and also encompass a similar amygdalar, affective component to that hypothesized with run time (McDonald et al., 2004; McDonald & Hong, 2004). The R1 trained population engaged in this behavior significantly more than control animals, both overall and independently of difficulty level (figures 7e-f, one-way ANOVA with Tukey HSD post-hoc, $F(1, 1414) = 72.1$, $p = 0.001$; supplementary figure S.3b, two-way ANOVA, no significant interaction of 'Group*Difficulty'). We also observed that both populations, independently and pooled together, performed significantly more choice revision on complexity level 0 trials compared to all other complexity levels, with in fact no significant difference in amount of choice revision behavior observed between the other complexity levels

(supplementary figure S.3b; only pooled results shown; pairwise t-tests within 'Difficulty' between trial complexity level 0 and each other level, $t(58) > 3.9$, p-values < 0.0002; between all other levels, $t(58) < 1.4$, p-values > 0.17). Curiously, however, we also observed that, looking specifically at choice revision towards S1 arms in the R1 trained population, choice revision on complexity level 1-4 trials was just as high as on complexity level 0 ones (supplementary figure S.3c), indicating an active cognitive deliberation-provoking interference from R1 in trials of all complexity levels. Choice revision decreased significantly over time/repeated EdRR training in the R1 trained population only, having started out in EdRR 1 at a much higher level than controls (repeated measures ANOVA within 'Session', R1 trained group, $F(11, 429) = 9.2$, corrected $p < 0.0001$; control group, $F(11, 198) = 0.7$, $p = 0.74$). In control animals, what little of this behavior we observed was equally balanced between the two surfaces but did occur more often in a choice rectifying rather than error-inducing manner (figure 6e, one-way ANOVA with pairwise Tukey HSD post-hoc; within control group, between correct/incorrect outcome, $F(1, 454) = 68.8$, $p = 0.001$). The R1 trained population, in contrast, revised their choice significantly more often towards S1 than towards S0 arms (one-way ANOVA with Tukey HSD post-hoc; within R1 trained group, between 'Surface', $F(1, 958) = 56.1$, $p = 0.001$). This effect was especially significant in EdRR 1 (figure 6e, pairwise t-tests with Bonferroni correction and unbiased Cohen effect sizes; EdRR 1, $t(78) = 4.1$, $p = 0.001$, $d = 0.91$; difference in EdRR 2, which appears significant on graph, was not significant following Bonferroni correction, $t(78) = 2.7$, $p = 0.09$). However, as seen in figure 6f, in EdRR 1 and 2 choice revision in the R1 trained population was equally likely to lead to an error as to a correction. It was only from EdRR 6 onwards that significantly more R1 trained population choice revision was rectifying rather than error-inducing (pairwise t-tests with Bonferroni correction and unbiased Cohen effect sizes; EdRRs 1-5 + 11, $t(78) < 2.5$, $p > 0.05$; EdRRs 6-10 + 12, $t(78) > 3$, $p < 0.05$), but even during these sessions, the overall significant trend for more choice revision to terminate with an S1 choice persisted (figure 6e, one-way ANOVA with Tukey HSD post-hoc on EdRRs 6 – 12, $F(1, 558) = 20.1$, $p = 0.001$), and this trend did not switch at any point during the 12 sessions of EdRR. Finally, since we had already observed in supplementary figure S.3b that significantly more choice revision occurred on complexity level 0 trials, and also that EdM performance was significantly higher at complexity level 0 trials (figure 4a, top),

we hypothesized that there would be an observable interaction in the data between these two results. Hence, we isolated complexity level 0 trials and separated them into two groups; those level 0 trials where no choice revision occurred ('KOOKs = 0') and those where at least one choice revision occurred ('KOOKs > 0') (figure 6g). We initially looked at block 1 in order to capture the period of strongest effects of the previous R1 learning. During this first block, the R1 trained population engaged in choice revision behavior on 81 out of 480 (16.9%) level 0 trials, and controls on 21 out of 228 (9%). What we observed (figure 6g) is that the EdM performance of the two populations was similar on level 0 trials when no choice revision occurred, but that only controls committed significantly fewer EdM errors on those level 0 trials where they engaged in choice revision behavior, as opposed to the R1 trained population who were just as likely to make an error even with choice revision (one-way ANOVA with Tukey HSD post-hoc, within each population between 'KOOKs = 0' and 'KOOKs > 0'; R1 trained population, $F(1, 173) = 0.05$, $p = 0.83$; control population, $F(1, 72) = 13.1$, $p = 0.001$). To be precise, in the case of the control animals, out of 21 trials on which choice revision occurred, only 1 of these gave rise to an error. With respect to surface, in the R1 trained population, 87% of the errors (102/117) made during block 1 on level 0 trials with no choice revision behavior ('KOOKs = 0') were towards S1, compared to 51% (35/68) in controls. On trials *with* choice revision ('KOOKs > 0'), in the R1 trained population 100% of the errors (26/26) were towards S1. As mentioned, in control animals only 1 error arose at this level when choice revision occurred, precluding any meaningful analysis with respect to distribution of errors by surface. These differences between the populations in choice revision behavior were most marked in block 1, as we had expected, but did also hold when analyzed across all training blocks averaged together, even though with repeated EdRR training the R1 trained population also began to make less error-inducing choice revision (supplementary figure S.3d, one-way ANOVA with Tukey HSD post-hoc, within each population between 'KOOKs = 0' and 'KOOKs > 0'; R1 trained population, $F(1, 661) = 3.8$, $p = 0.051$; control population, $F(1, 282) = 14.4$, $p = 0.001$). Averaged across all blocks, 236/316 (75%) of all level 0 errors without choice revision in the R1 trained population were S1 choices, while 38/43 (88%) of all level 0 errors *with* choice revision were. Taken together, these analyses of choice revision behavior confirmed what we had predicted: that increased cognitive deliberation was just as, if not more, likely to give rise to R1 biased choice behavior as less deliberatively

made choices were. In short, R1 bias, in all the forms we observed it during EdRR, was no simple affair of unthinking persistence of previous learning.

6. Intra-session, surface independent choice-confirmation bias also contributes to pair-by-pair EdRR errors.

Beyond the myside bias effect characterized above, we had also predicted that we may observe a more immediate, proximal intra-session confirmation bias effect. What we mean by this is an increased probability on a given trial *n* to try to confirm the outcome of the previous trial on the same pair, *n-1*, when that outcome was rewarded as opposed to not rewarded. This type of more proximal confirmation bias is a cognitive phenomenon which has been the object of much recent human and computational investigation under the label of "choice-confirmation bias" (Chierchia et al., 2021; Palminteri, 2021; Palminteri et al., 2017). Thus, although our experimental conditions were designed specifically to elicit myside confirmation bias-like effects, they did also offer the opportunity to measure choice-confirmation bias-like effects, if these were present. Hence, looking at trial-by-trial performance as a function of the previous outcome on a given pair ('Pair N-1'), we observed that both the R1 trained and control populations were significantly more likely to choose the same surface again if the outcome on the previous presentation of the current pair had been a reward. In other words, they were more likely to try to *confirm* a previous outcome in the case that this had been rewarded. In order to disentangle, as much as possible, any such proximal intra-session confirmation bias effect from the R1 bias effect, we conducted analysis of S1 and S0 outcomes independently (figure 7a-b). Firstly, with respect to surface choices following either S1 rewarded (S1+) or S1 unrewarded (S1-) trials, we found that both R1 trained animals and controls were significantly more likely to choose S1 on trial *n* when *n-1* had been an S1+ rather than an S1- outcome (two-way ANOVA between 'Pair *n-1* outcome' and between 'Group', averaged over all blocks of EdRR; highly significant effect of both 'Pair *n-1* outcome' and 'Group', $F(1, 1401) = 164$, $p < 0.0001$, $F(1, 1401) = 39$, $p < 0.0001$, respectively).

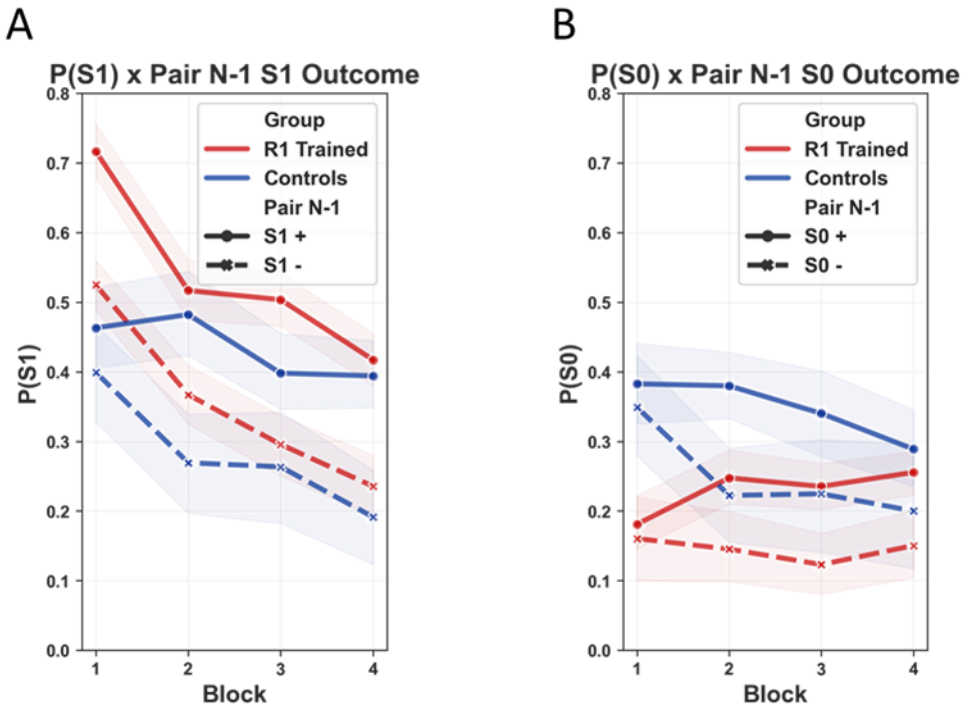


Figure 12 - Intra-session choice-confirmation bias observed in both R1 trained and control populations.

Both controls and, despite their previous S1-reward association learning, R1 trained animals displayed choice-confirmation bias-like behavior relative to both S1 *and* S0, which we analyzed separately precisely to disentangle the results from the R1 bias. (**A**) P(S1). Analyzing at the level of 3 session blocks, we saw that both R1 trained and control populations were significantly more likely to choose S1 on a presentation *n* of a pair when the outcome from the presentation *n-1* of the same pair had been S1+ (i.e. rewarded) rather than S1- (i.e. not rewarded). The trend was present in all blocks, but did not reach statistical significance only in controls and only in block 1. (**B**) P(S0). Similarly both the R1 trained and control populations were significantly more likely to choose S0 on a presentation *n* of a pair when the outcome from the presentation *n-1* of the same pair had been S0+ rather than S0-. The trend was present in all blocks, but notably did not reach significance in either population in block 1. The reduced effect sizes in block 1 could, pending further investigation, be the anomalies here, especially when it is recalled that block 1 englobes all the novelty effects of the transfer from the R1 to the EdRR environment.

However, the same statistical analysis revealed no significant effect of the interaction 'Pair *n-1* outcome*Group' ($F(1, 1410) = 1$, $p = 0.3$). These statistical results indicate, as figure 7a suggests, that although there was an overall 'between Group' effect (certainly due to R1 trained animals choosing S1 significantly more than controls over), there was no significant difference between the effect sizes of their respective 'between Pair *n-1* outcome' differences. More convincingly still however, was the fact that both the control and the R1 trained populations also displayed a higher probability of choosing S0 on a trial *n* if the outcome on trial *n-1* had been S0+ rather than S0-. While the effects seen with respect to S1+/S1- may to some extent be related to previous learning in the R1 trained population, or to P(Rw|S1) being higher than P(Rw|S0) in controls, that explanation would not account for our observation of the same effect with respect to S0+/S0-. These results open the possibility that this choice-confirmation bias-like effect is a parallel cognitive mechanism which can nevertheless interact with myside-confirmation bias-like effects in such a way as to amplify their manifestation. By the facts of the EdRR rule, choosing, for example, S1 following either an S1+ *or* an S1- will necessarily give rise to an error and therefore to an absence of reward, meaning that whatever inherent bias there is towards an effect such as this, it may be partially restrained by the EdRR experimental design. In a scenario where errors were not rapidly guaranteed, it is possible this choice-confirmation bias-like effect would be stronger.

Discussion

The present study is the first to provide robust evidence of non-human animals manifesting myside confirmation bias-like behavior in situations where a previously internalized behavioral rule/state-action policy (Sutton & Barto, 2018) must be revised in the context of a novel environment which is semantically ambiguous with respect to it, in the sense that information from this novel environment sometimes confirms and sometimes disconfirms the prior rule. The results constitute validation of a mouse model of real-world, everyday-like human situations in which previously reinforced beliefs, understood as action policies, need to be updated or revised in light of novel information coming from a dynamic and complex environment. Beyond its inherent interest in providing the means to gaining a deeper understanding of how the mammalian mind-brain cognitively and neurobiologically adapts to such situations, the introduction of

this model opens many avenues for future pre-clinical research in domains where differential confirmation bias-like phenotypes have been observed in humans, such as addiction rehabilitation and behavior change (Granero et al., 2020; Prochaska, 2008), schizophrenia (Doll et al., 2014), ageing (Wilson et al., 2018), as well as other physiological (Rollwage et al., 2020; Rollwage & Fleming, 2021; K. Stanovich et al., 2013; K. Stanovich & West, 2007) and pathological (Balzan et al., 2013) conditions. Furthermore, this model validation in and of itself alters the research environment with respect to our current thinking around the neurobiological and environmental evolution of the higher reasoning faculties of humankind. What follows is a comprehensive discussion of our results, their implications, and what they suggest for future research directions.

Impact on everyday-like memory performance of a previously acquired, partially antagonistic cognitive rule.

When animals "indoctrinated" to reliably respond to the tactile stimulus-response rule (R1) (Stevens et al., 2022a) were transferred to the EdRR environment, our first observation was that this prior learning generated only a slight although initially significant negative impact on overall EdM performance compared to controls, whose prior learning experience was to have been rewarded on every trial. An important cognitive consequence of this control condition is that both R1 trained and control animals were dealing, in the EdRR phase, with comparably radical, albeit distinct, changes to the response-outcome (R-O) reward contingencies they had become accustomed to over repeated exposures to the R1 radial maze environment. This fact of our experimental design entailed an important advantage in interpretation of the results, in that it provided us with a control for the initial frustration the R1 trained population was likely to experience in the EdRR environment upon not obtaining reward when expected (Amsel, 1958; Martín-García et al., 2015), an interesting phenomenon in itself, but not the direct object of our investigations. The necessity to adapt to such radical changes in reward contingency may indeed explain why EdM performance in the EdRR environment was much lower, even in controls, than performances normally observed in the classical EdM task with naïve animals (Marighetto et al., 2011), and it is important that this point be underlined: overall EdM performance *does* appear to be

negatively affected by R1 training, *and also* by our R1 phase control condition. Nevertheless, one may have expected the initial mean EdM performance in the R1 trained population to fall below chance level, which indeed would have been the case had they chosen S1 arms during EdRR 1 according to their mean likelihood of choosing S1 during the final R1 session, i.e. 83%. Why did this not happen?

EdM performance and the impact of novelty on exploration.

The links between novelty and exploration in both humans and rodents are well established and have been the subject of much research and discussion (Farahbakhsh & Siciliano, 2021; Lustberg et al., 2020; McDonald et al., 2004; Park et al., 2021). The most significant observation from our experiments in this regard is that, upon initial introduction to the novel EdRR environment, which we had expressly made as sensorially novel as possible (new room, new radial maze, new lighting conditions, new ambient odor), not only did the R1 trained population not immediately attempt to explore S1 arms, they pointedly did the opposite. A majority of R1 trained animals chose the S0 surface on the very first EdRR 1 trial, followed by an even greater majority of the remainder on the second trial. In fact, a significant majority of first trial S0 choosing animals also chose S0 on the second trial despite having received no reward from their first S0 choice (recall the optimal policy for performing EdRR; choose S1 on initial per session presentation of each pair, then spatially alternate on all subsequent pair presentations). This result was as surprising as it is fascinating, since it indicates that environmental novelty can drive not only topographical exploration of the new environment in mice, but also *nomological* exploration directed precisely beyond the bounds defined by an internalized behavioral rule. The possibility can even be suggested that these two modes of exploration are connected, potentially subsumed under an information gain exploratory drive (Inglis et al., 2001), a question we intend to explore in future work as part of a computational modeling approach to deeper understanding of the behaviors observed in mice in this study (also discussed in (Stevens et al., 2022a)). Recalling that, for control animals, absence of reward on an initial S0 choice during EdRR constituted the very first time they had encountered a no-reward outcome on a trial, thus beyond environmental novelty, there was also a putative effect of frustration or surprise at play (Amsel, 1958; Xu et al., 2021). This could explain why, of those

controls who chose the rewarded S1 on the first trial of EdRR 1, only half (i.e. commensurate with random choice behavior) chose S0 on the second trial, whereas of those who chose the unrewarded S0 on the first trial, a majority chose S0 again on the second trial, revealing a higher probability for a counter-intuitive "lose-stay" strategy, which we do not believe can be satisfactorily explained by decision inertia (Alós-Ferrer et al., 2016) and therefore merits further investigation.

Following on from this last point, classical studies on the various manifestations of confirmation bias in humans have focused heavily on the finding that human subjects primarily engage in rule testing via confirmation rather than refutation, despite the fact that in terms of information gain the latter strategy has the capacity to return more decisive information. If we consider the initial trials in EdRR 1 as a putative animal analog of one of Wason's classical human reasoning tasks from the psychology literature (Wason, 1960, 1966, 1968), it is curious to remark that the majority of R1 trained mice began by making refutational rather than confirmational action decisions with respect to the R1 rule, i.e. by first choosing S0 rather than S1. Above chance-level probability of R1 trained populations choosing S0 on the first EdRR trial was highly replicable across all iterations of the protocol we conducted, including those with transgenic and aged mice (data not shown). During final R1 sessions also, we had observed that the R1 trained population was more likely to first explore an S0 arm early rather than later in the session, earlier even than could be predicted on the basis of their S1 choice probability from the previous session (Stevens et al., 2022a). This apparently refutational behavioral was less marked in R1 sessions than in EdRR 1, but taken together both invite the intriguing possibility that laboratory mice will more spontaneously attempt to refute a rule than adult humans will. This eventuality in turn suggests that the specific bias towards rule confirmation identified by Wason may be at least in part more a result of education than nature, a suggestion already put forward by Raymond Nickerson in his classic literature review on confirmation bias (Nickerson, 1998) and seemingly backed up by recent studies showing that "win-stay"-like confirmatory behavior increases in humans as a function of maturation (Chierchia et al., 2021).

Manifestation of the novelty-induced exploratory drive was not limited to only the first two trials of EdRR 1. The median number of trials taken by the R1 trained population

to at least once explore all six available arms, S1 and S0, of the EdRR radial maze approached optimal level (i.e. 7 trials) during EdRR 1, making them stronger initial explorers than even controls. Since we hypothesized that this burst of exploration was driven primarily by the salient novelty of the EdRR environment, we also predicted that we would see less striking exploratory behavior later in EdRR 1 and also in EdRR 2. This prediction was borne out by the data. First, on all trials from 3 to 22 of EdRR 1, the R1 trained population were more likely to choose S1, with the surface choice probability shifting towards S0 only on the final 23rd trial, tellingly a complexity level 0 trial. Subsequently, we observed that the median exploration score in the R1 trained population increased significantly in EdRR 2, i.e. mice took longer to explore all available arms at least once during EdRR 2 compared to EdRR 1, fitting with our hypothesis that the initial burst of exploration was primarily a cognitive response to novelty.

As mentioned above, there is some precedent in the literature demonstrating a facilitation of S-R reversal learning when this takes place in a novel context rather than in the same context as the initial learning (McDonald et al., 2004). However, our EdRR protocol was designed precisely not to be a simple reversal paradigm, and while previous results in rule reversal follow a steady upwards curve, here we notably saw no mean difference in P(S1) in experimental animals between EdRR 1 and EdRR 2, despite an average of ~60% of all EdRR 1 trials giving rise to disconfirmations of R1. Thus, novelty, rather than repeated errors from behaving according to R1, was the primary driver of initial exploratory, i.e. S0, choice behavior. In the language of reinforcement learning then, what these results with regards to exploration during EdRR 1 and EdRR 2 seem to reveal is that the R1 trained population had no deficit compared to controls with respect to their initial exploratory sampling of a novel environment, but they were impaired in subsequently using their experiences of R1 disconfirmation to appropriately update their level of adherence to the R1 rule. Finally, if we compare this to other recent work which demonstrated, in naïve mice, that increased novelty improves consolidation of proximal experiences (Takeuchi et al., 2016), this gives an even better idea of just how strong the impact of myside confirmation bias-like effects from prior learning can be, since the manifestation of this bias which we observed seemed precisely to relate to a failure to sufficiently consolidate R1 disconfirmations from session to session.

Win-stay versus win-shift strategies.

The primary finding from our analysis of the indoctrination-like R1 training phase of our protocol was that, in maze conditions, win-shift behavior comes more spontaneously to mice than win-stay and that the transition to the latter is mediated in part by local modulatory inhibitory control (i.e. via the retrograde inhibitory action of CB_1 receptors) over the direct pathway of the dorsal striatum (Stevens et al., 2022a). Indeed, we advanced that R1 expression required ongoing inhibition of the spontaneous exploratory drive, and this was corroborated by our EdRR investigation in two ways. The first we have just seen; initial environmental novelty is sufficient to lift inhibition of exploration and, despite their intense win-stay training, R1 trained mice demonstrate that their exploratory drive, when not inhibited, is as active as ever. Secondly, when local modulatory inhibitory control in the direct pathway was reduced, by conditional deletion of the CB_1 receptor from D_1 expressing neurons of the forebrain (see Appendix), a certain number of these transgenic animals did not reach criterion R1 performance. Based on our inhibitory control hypothesis, we suggested that this was due to a reduced capacity to inhibit the direct pathway-mediated win-shift exploratory drive. And indeed, subsequently when we moved these below R1 criterion D_1-CB_1-KO mice to the EdRR environment, we were able to observe that they instantly and persistently performed significantly better on the win-shift based EdRR task than their above R1 criterion D_1-CB_1-KO littermates. This convincingly demonstrates that learning which demands inhibition of spontaneous exploratory behavior is a central component of subsequent myside confirmation bias. Nevertheless, this same below R1 criterion D_1-CB_1-KO population, who committed significantly fewer EdRR errors and displayed lower overall P(S1) than their wildtype littermates, were just as biased towards S1 on those errors that they did commit, showing that they were still impacted, albeit less so, by the indoctrination-like schedule of the R1 training phase.

Again in the R1 training phase, we corroborated our hypothesis regarding the role of local modulatory inhibitory control in the direct pathway by virally re-expressing CB1 receptors on the D1-expressing neurons, which make up the direct pathway of the dorsal striatum, of D_1-CB_1-KO mice. This population (D_1-CB_1-KO-$Str_{+/+}$) displayed a significant facility in attaining and maintaining criterion R1 performance, which is to say, in

inhibiting spontaneous exploration, during the R1 phase (Stevens et al., 2022a). Relative to our current study, these same enhanced R1 expressing animals did not however perform worse in EdRR nor display more bias than D_1-CB_1-KO-$Str^{-/-}$ littermates, injected with an empty viral vector. This may in itself be a reflection of just how much EdRR performance relies on the interaction of multiple systems, as opposed to R1 performance which is predominantly striatal (Stevens et al., 2022a).

Finally, the observation that decision latency increases as a function of successful performance of R1, compared to controls, but then decreases to be equivalent to control levels as a function of repeated exposure to the EdRR task, is another indication that the cognition required for R1 is not spontaneous, whereas that required for EdM is. Thus, although the R1 rule may seem formally simpler from a human perspective, it seems clear that from a murine point of view, EdM responding comes significantly more naturally. However, in the particular case of EdRR, neither responding rapidly nor, as we have seen in the case of choice revision, are any guarantee of protection from the expression of bias from previous learning.

Within-error R1 bias.

While only a slight and initial significant difference was observed between the R1 trained and control populations in the overall number of session-by-session EdM errors, the nature of the errors committed by the two groups was vastly different. On average, the R1 trained population committed almost 90% of their errors on S1 arms in EdRR 1 compared to an almost even split of errors across S1 and S0 arms observed in controls. To obtain a more robust evaluation of the extent of R1 error bias manifest in the EdM errors, we calculated the relative difference between error types (S1 errors – S0 errors/Total S1 + S0 errors) to produce a within-error R1 bias index. Curiously, however, we found no correlations between R1 training phase performances (either final or summed or weighted) and R1 error bias, nor even between EdRR 1 exploration score and R1 error bias, all of which is further indication of the multiplicity of cognitive processes mobilized by our EdRR protocol, e.g., respectively, the cognitive mechanisms necessary for inhibiting the innate exploratory drive during R1 training and those involved in sensitivity to environmental novelty and inhibition of interfering/intrusive mental content, etc. This was the first of many reflections in our mouse model of the

finding in human studies that myside bias is not correlated to measures of general intelligence (K. Stanovich et al., 2013; K. Stanovich & West, 2007). By the final, fourth block of EdRR training (sessions 10-12), the within-error R1 bias index was no longer different between the R1 trained and control groups, with both groups displaying a slight within-error R1 bias. Naturally, this raised the question of why the control group may have developed such a bias during EdRR training. One potential explanation is the contingent fact that the overall session-by-session probability of receiving a reward having chosen an S1 arm, i.e. P(Rw|S1), was reliably higher than P(Rw|S0) in controls across all EdRR sessions. Thus, some kind of putatively Bayesian and striatal statistical process of updating R-O values from session to session could explain this trend in controls (Kim et al., 2009; Samejima et al., 2005). In the case of the R1 trained population, by contrast, P(Rw|S1) was significantly lower than P(Rw|S0) until the fourth block of EdRR, and thus their within-error R1 bias persisted not because of but *despite* these experiences.

R1 bias as a function of trial complexity.

As discussed above, only a slight significant overall difference was observed in EdM performance between the R1 trained and control populations. This difference, present during the first block only, was driven by the R1 trained population making significantly more EdM errors than controls on the most complex trials only (levels 3 and 4). During the first block of EdRR, performance in controls did not drop below chance level at any trial complexity. Here and in previous studies (Marighetto et al., 2011), chance level performance in the EdM task has been interpreted as a reflection of animals not being able to accurately recall their choice from the previous presentation of the current pair, thus leading them to choose randomly the arm to visit on the current trial, thereby giving rise to the chance level performance. Also according to this interpretation, animals perform more EdM errors on more complex trials because these trials imply not only more interposed, interfering cognitive content, but also more "passive" forgetting with the passage of time, all of which creates a perfect cognitive storm for clear recollection of the relevant past action, i.e. which arm was chosen on the previous presentation of the current pair. In terms of R1 interference in EdM performance during EdRR, an interesting parallel emerges in the observation that, in the R1 trained population, the

probability of choosing S1 ('P(S1)') also increased significantly as a function of trial complexity. Thus, it seems that the probability of an R1 trained animal choosing S1 on the most complex trials was so high during the first block that this actually pushed their total number of errors beyond chance level, the level at which controls were performing on those same complexity level trials.

An affective contribution to biased cognition?

Myside bias in humans is commonly referred to as the irrational, emotional dimension of human reasoning (Pinker, 2021; K. E. Stanovich, 2021). In (Stevens et al., 2022a) we already advanced that the marked differences observed between run times on S1 and S0 arms could, based on previous studies, have an affective, putatively amygdalar component (McDonald et al., 2004; McDonald & Hong, 2004). During early EdRR sessions, we observed that the higher run times associated with S0 decreased only gradually over repeated training. Furthermore, lower run times on S1 were observed up until the 12th session of EdRR, indicating some level of highly persistent "over-confidence" in S1 choices even after 9 sessions of training during which P(Rw|S1) was reliably lower than P(Rw|S0). This is also a clear demonstration that bias can and does manifest beyond instances of error.

Similarly, since in the EdM and EdRR tasks, mice have the possibility of advancing along an arm prior to revising their choice (provided that they do not reach the final distal zone), there may also be an affective contribution to the significantly S1 favoring choice revision behavior we observed.

Every episode of choice revision during EdRR could constitute an almost literal window into myside-like bias as it is happening: initial cortico-hippocampal cognitive decision making impacted after the fact by striatal statistical and amygdalar affective inputs which bias the response towards the previously internalized rule. During the 12 sessions of the EdRR phase with young adult C57Bl6/J mice, we identified 1,275 choice revision episodes in the R1 trained population, making of such events robust and fertile ground for future investigations and interventions.

Myside confirmation bias: failure to cognitively identify and inhibit intrusive thoughts?
Intrusive thoughts and their inhibition through a cognitive mechanism referred to as active (or adaptive) forgetting have been the focus of much recent investigation (Anderson & Floresco, 2021; Anderson & Hulbert, 2021; Bekinschtein et al., 2018; Costanzi et al., 2021; Geraerts et al., 2007). While most of the literature on active forgetting has thus far concentrated on cortico-hippocampal inhibition, it is worth noting in the context of our EdRR task that medial prefrontal cortex (mPFC) projections to the dorsal striatum also exert inhibitory control in rodents (Terra et al., 2020). One of the particularities of the EdM and EdRR behavioral paradigms is the fact that, before they can be inhibited, those thoughts which are intrusive relevant to the present trial must first be contextually identified from interactions with the environment. This process of identification notably becomes more difficult the more complex (in terms of number of interposed trials on other pairs) the present trial is. So while we observed that on level 0 trials R1 trained animals had a high level of success in exerting inhibitory control over the R1 striatal strategy, this capacity decreased as trial complexity and thereby response uncertainty increased. With regards to increasing uncertainty as trial complexity increases in the EdM and EdRR tasks, light may be shed on this by research demonstrating that active inhibition of cognitive content can provoke a kind of amnesia with respect to it (Hu et al., 2017; Hulbert et al., 2016). Thus, on a complexity level 4 trial, for example, the theory of adaptive forgetting suggests that the memory of the response to the last presentation of the present pair has been actively inhibited four times, i.e. on each of the four interposed trials where that memory was not relevant. In this sense, we can see that it is a fine line between adaptive and maladaptive forgetting.

Furthermore, in this regard, in forthcoming work from our team, we have found that poor mnemonic performance in aged mice on EdM tasks with a high organizational demand correlates with increased activity in the hippocampus compared to young adult mice. For this reason, we have begun to theorize the possibility that at least one dimension of age-related mnemonic cognitive decline is, counter-intuitively, not an *excess* of forgetting, but rather a *deficit* of adaptive forgetting.

Perhaps most promisingly with regards to future research avenues, through our experiments in the EdM and EdRR tasks, we were able to demonstrate, for the first time, a central role for cannabinoid type-1 receptors expressed on GABAergic neurons of the

forebrain in adaptive forgetting. Indeed, in the Dlx-CB_1-KO population we seem to have discovered a phenotype of near total incapacity for active forgetting, despite this mouse line having no deficit in basic (i.e. without a notable inhibitory component) spatial working memory (Albayram et al., 2016) or spatial retention memory (Han et al., 2012). In preliminary work we conducted in the classical EdM task (see Appendix), Dlx-CB_1-KO reliably performed worse on complexity level 0 trials (which closest resemble a basic spatial alternation task) than both aged mice and mice with lesions of the CA1 (Marighetto et al., 2011). This was further confirmed by their performance in the present EdRR task, where unlike all other R1 trained populations, including aged mice, Dlx-CB_1-KO did not manage to achieve better than chance performance even on level 0 trials. In fact, under the weight of the R1 bias, they actually performed significantly lower than chance on level 0 trials in the first block of EdRR sessions. However, looking especially at the classical EdM task, we also observed significantly higher pre- and post-choice deliberative behavior (decision latency, run time, choice revision) in Dlx-CB_1-KO mice compared to wildtype, strongly suggesting a surplus of active cognitive content, which is what we should expect to see where there is a dysfunction of active forgetting. Taken together, this is very strong evidence that inhibitory control requires the population of CB_1 receptors expressed on GABAergic neurons of the forebrain, and thus if inhibitory control is fundamentally linked to real world working memory demands, then working memory does also require GABA-CB_1. A phenotypical and putatively functional similarity between aged and Dlx-CB_1-KO mice has been previously identified in a Morris water-maze spatial learning task (Albayram et al., 2011). Thus, echoing recent research showing that chronic THC administration improves spatial and reversal memory in aged mice (Bilkei-Gorzo et al., 2017), our results confirm the potential of the endocannabinoid system as a therapeutic target for age-related memory decline, with much pre-clinical research yet waiting to be conducted.

Finally on this point, during discussion of our results with other researchers, one recurring comment was that, since R1 trained animals were apparently not biased on level 0 trials, then they could not be said to have a strong myside confirmation bias. However, we believe that the capacity for mPFC inhibition of striatal responses in simple contexts can explain this, in mice and potentially in humans also. Indeed, context inappropriate thoughts, as we have already discussed, are precisely easier to identify and inhibit in simple contexts. Consider two toy examples of situations where the correct

response would be evident, even to someone who may otherwise be committed to doctrines that give rise to strong confirmation bias in situations of higher inherent uncertainty: I've just broken my arm; will I go to the hospital or will I seek a homeopathic remedy? I want to know what the weather is going to be like tomorrow; will I consult the meteorological forecast or my horoscope? Intrusive thoughts may simply no longer be recognized as such once task-inherent uncertainty rises above a certain noise-to-signal ratio. Indeed, we even observed that R1 trained animals who hesitated about and ultimately revised their initial choice on level 0 trials were just as likely to make a (biased) error than when they did not revise their choice, whereas choice revision in controls was almost always rectifying at level 0. And this is despite the fact that when R1 trained animals did *not* revise their choice on level 0 trials, they performed as strongly as controls, indicating that, in the case of a biased agent, opening the door to uncertainty, e.g. by hesitating, is widening the point of entry for bias into the decision making process. Again, as an observable behavior this corresponds closely with a certain interpretation of human reasoning reached on the basis of myside bias, i.e. that we spend more of our cognitive energy on rational*izing* than we do on *being rational* (Mercier & Sperber, 2017). In brief, for mice in the EdRR task as for humans in everyday life, simply "thinking more" offers no protection against bias. This again also relates to the decoupling of levels of general intelligence from susceptibility to myside bias observed in humans (K. Stanovich et al., 2013; K. Stanovich & West, 2007).

In certain contrast to mice during EdRR, however, humans do possess a powerful capacity for outcome reinterpretation. We do not have the scope to develop this point here, but the cognitive mechanism referred to informally as "sour grapes", after the fable of the fox and the grapes from Aesop and LaFontaine, precisely allows for absence of reward to be reinterpreted as a positive event, often making this the most "rewarding", globally "reinforcing" cognitive option by which to "protect" the integrity of established neural pathway circuitry (Kaplan et al., 2016). Thus, this manner of resolving cognitive dissonance (Festinger, 1957) may certainly contribute to myside bias in humans, but is in fact a distinct cognitive function.

Choice-confirmation bias-like effects

Related to the human capacity for reinterpreting outcomes, we will very briefly return to the preliminary investigation we conducted with respect to *choice*-confirmation bias-like effects, i.e. a higher likelihood to try to *confirm* a choice when it led to a positive rather than to a negative outcome. As we demonstrated, we were able to measure just such an effect, to comparable extents, in both the R1 trained and the control populations, with respect to both P(S1) being higher following S1+ versus S1- outcomes and P(S0) being higher following S0+ versus S0- outcomes. On this point, it is also worth noting that by the facts of the EdRR rule, choosing, for example, S1 following either an S1+ *or* an S1- will necessarily give rise to an error and therefore to an absence of reward, meaning that whatever inherent kind of choice-confirmation bias there may be in mice, it may actually be partially restrained by the EdRR experimental design. In a scenario where errors were not so rapidly guaranteed, it is possible this choice-confirmation bias-like effect would be stronger. In humans, the potential for outcome re-interpretation, a mechanism of resolution of cognitive dissonance (Festinger, 1957), may thus in turn also amplify choice-confirmation bias-like behaviors in circumstances where, for example, we convince ourselves our *choice* was good even if the *outcome* wasn't. The question of the precise cognitive, neurobiological, and evolutionary roots underpinning such interactions between distinct behavioral and epistemic mechanisms remains rich terrain for future investigation in both humans and animals.

Majority of bias suppression occurs through striatal and/or affective functions.

As we have just seen, both aged and especially Dlx-CB_1-KO mice are impaired in their capacity for active forgetting. Interestingly, however, extinction of the R1 bias proceeded at a similar rate in both of these populations as compared to wildtype littermates and all other populations tested. In the case of Dlx-CB_1-KO, they did display slightly more R1 bias than wildtypes in each session, but declining at a comparable rate. Once again reinforcing the idea that higher cognitive ability is largely independent of myside-like bias, this shows that the cortico-hippocampal capacity to perform the EdRR task offers only limited protection from the effects of R1 bias. As demonstrated by their unaffected performance during R1 training (Stevens et al., 2022a), Dlx-CB_1-KO display normal striatal function. Therefore, if, as we have already suggested, the striatum is

responsible for incremental, statistical updating of response-outcome values, this, and not cortico-hippocampal function, could be the major contributing factor to extinction of R1 bias. In parallel, surface-based differences in run time and choice revision, which we have at least partially related to amygdalar/affective function, also declined normally over repeated EdRR training in both aged and Dlx-CB_1-KO mice.

A multiple memory systems conception of myside confirmation bias.

Taken together, the results validating the EdRR model of everyday-like rule revision and consequent myside confirmation bias, as well as the results from our preliminary transgenic and ageing approach interventions using the model, open up a compelling possibility: that the EdM and especially EdRR task are uniquely capable of differentially mobilizing up to four interacting (competing and/or cooperating) memory systems: episodic memory, striatal memory, affective memory, and working memory (englobing its dimension of inhibitory control; figure 8). Beginning with working memory, it is widely accepted that Y- and T-maze spatial alternation preferentially engages this memory system (Albayram et al., 2016; Aultman & Moghaddam, 2001; Jobson et al., 2021; Shoji et al., 2012). In the classical EdM task also, various studies from our team, including the present one, have shown that animals with either lesion of the CA1 region of the hippocampus (Marighetto et al., 2011) or genetically induced episodic memory dysfunction (i.e. Dlx-CB_1-KO, seen above) are still capable of above chance level performance, but in level 0 trials only (and in the case of Dlx-CB_1-KO only slightly above chance). These results suggest that hippocampal memory is necessary for level 1 to level 4 complexity trials but not for level 0 trials, a conclusion that also fits with well-established hypotheses which take working memory to be a cortical rather than hippocampus-centered memory system (Baddeley, 2003; Curtis & D'Esposito, 2003; Lara & Wallis, 2015). Nevertheless, with our results from Dlx-CB_1-KO and aged animals, we also saw that the hippocampus is not sufficient for success on level 1 to 4 trials; top-down cortico-hippocampal active inhibition of interfering cognitive content is also required.

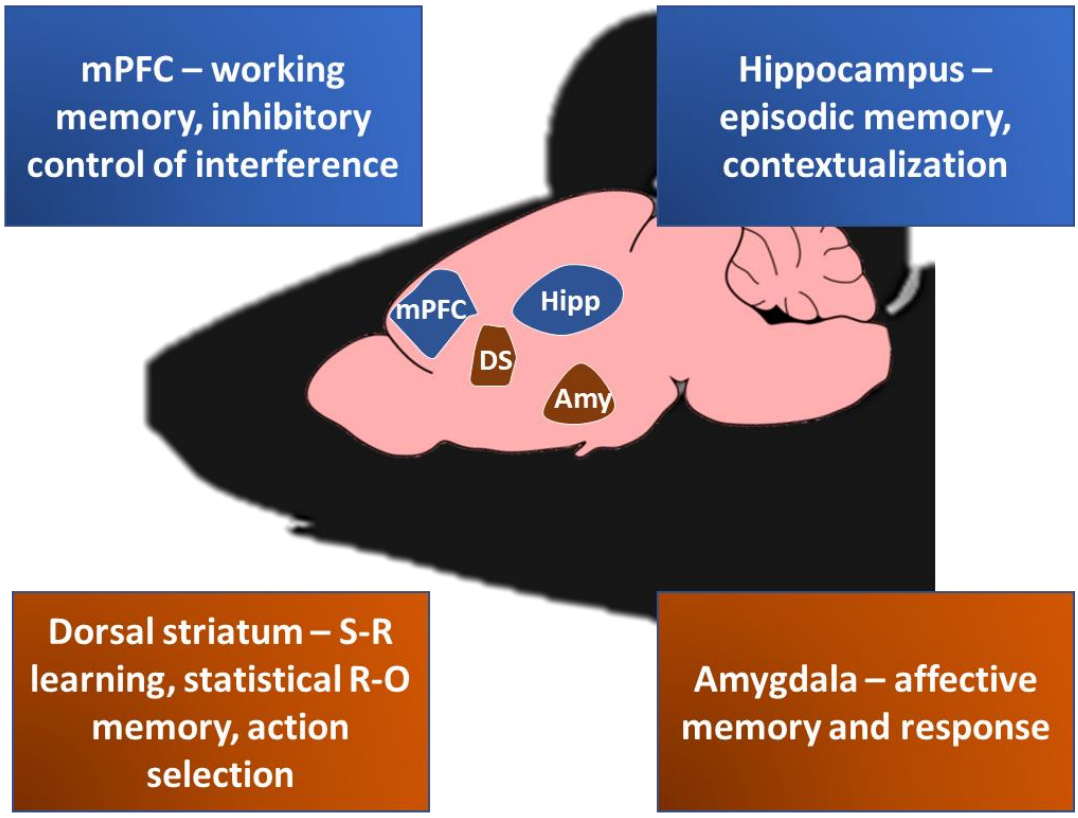


Figure 13 - Anatomical and functional mapping of multiple memory systems.

Translational interpretation of our results leads us to suggest that the cortico-hippocampal contribution (blue squares above) to EdRR behavior may be most closely related to what is measured under the label of general intelligence or cognitive ability in humans, which has been shown not to correlate with individual strength of myside side bias, just as we found no correlation between EdM performance and either strength of initial R1 expression during training or within-error R1 bias during EdRR, the primary contributors to which, we hypothesize, may be signals from deeper and more ancient structures, such as the striatum and the amygdala.

In our schema of these multiple memory systems (figure 8), we have colored the cortical and hippocampal systems in blue as being indicative of those memory systems most closely related to measures of general (semantic, associative, etc.) intelligence in humans.

Increasing complexity in the EdM task is a matter of more and more uncertainty being added to representational episodic memory recall via a combination of increased interfering cognitive content from a higher number of interposed trials (*or*, as proposed above, an amnesia effect produced by adaptively forgetting this content on trials where it is not needed) and time-dependent decline in the availability of precise and trial-relevant cognitive content. During EdRR, on trials of high complexity, we observed that the R1 trained population was significantly more likely to respond according to the S-R, striatally acquired R1 rule than according to the exploratory spatial alternation EdM rule. It is therefore plausible to suggest the following mnemonic model of the behavior; as contextualized cortical and hippocampal representational memory systems become noisier, due to increased uncertainty, so action choice will increasingly fall to decontextualized S-R striatal and/or affective amygdalar memory systems. In our schema (figure 8), we have colored the striatal and amygdalar systems in orange as being indicative of those memory systems most closely related to habitual, emotional, and therefore most often labelled "irrational" responding, typically not directly elicited by tests of general intelligence. If striatal R-O values are updating in some kind of ongoing Bayesian manner (Ballard et al., 2018; Kim et al., 2009; Markowitz et al., 2018; Nonomura et al., 2018; Samejima et al., 2005), incorporating the context-independent full historic of P(Rw|S1) from both the EdRR and prior R1 environments, and if the striatum is more heavily relied on to select action on more complex trials, then this could explain why P(S1) remains significantly higher on complex trials compared to easier ones until at least block 4 of EdRR. We can also bring some convergent evidence to this multiple memory systems hypothesis from our control group data. Here, we saw how, starting from EdRR 1, P(Rw|S1) was reliably higher than P(Rw|S0) in all sessions. In a breakdown of both P(Sn) and within-error R1 bias according to trial complexity, we also saw that controls developed a significant trend to choose S1 more than S0 on level 4 trials and also committed a significantly higher proportion of errors towards S1 on level 4 trials in blocks 2 to 4. If resorting to a striatal, and therefore statistical, response strategy on the most complex trials is an innate feature, then since, overall, control mice were statistically experiencing higher reward probability when they chose S1 compared

to S0, it is plausible to suggest that this statistical S-R response strategy is what we saw manifest in their significantly S1 biased level 4 surface choices. Similarly, it is not impossible that for the same reason (S1 more often rewarded than S0), control mice also developed a better "feeling" (i.e. via the amygdala) about that surface which could also have biased their choices in that direction in the absence of a clear cortico-hippocampal signal. However, we saw no sign of such an amygdalar contribution in control mice in either their run time or choice revision behaviors, where no surface-based phenotype developed over time, indicating to us (a question for further investigation) that the striatum is more sensitive to subtle differences in response-outcome values.

If such an ongoing surface-based statistical evaluation is indeed taking place during the EdRR phase, then rule revision with respect to R1 may be a case of taking one step backward for every two steps towards overcoming R1 responding in favor of the spatial alternation behavior necessary for successful EdM performance. This would seriously delay the updating process compared to, for example, more classical all-or-nothing reversal learning protocols, and could therefore be a key component in the lag we observed in R1 trained animals with respect to their updating of P(Sn) as a function of both P(Rw|Sn) – whereby P(Rw|S0) was significantly higher than P(Rw|S1) for the first 9 sessions of EdRR but P(S1) remained significantly higher than P(S0) throughout all 12 sessions – and the balance of R1 confirmations versus disconfirmations – where per session R1 disconfirmations significantly outweighed confirmations for the first 5 sessions of EdRR.

Finally, striatal memory would not play a role *only* in the most complex trials. Since the striatum is essential to the selection and initiation of movement, it is ultimately involved in decision-making, and especially decision execution, at all levels of complexity. Indeed, in the R1 trained population, striatal S-R responding leads to significant interference at all trial complexity levels, but at levels 3 and 4, striatal response may be the only recourse left to the organism. The role of the striatum in action choice has previously been modelled as a softmax action selection rule (Kim et al., 2009) whereby the probability of choosing one action over another (e.g. "explore arm on the left" vs "explore arm on the right", or "respond to S1" vs "respond to S0") varies according to the difference between the values attributed to each of the potential actions (e.g. $a = (Q_{S1} - Q_{S0})$, where a stands for the chosen action and Q stands for the estimated/predicted value of each

action option (e.g. Q_{S1} = estimated/predicted value of choosing S1), as per the *Q*-Learning reinforcement learning approach (Watkins & Dayan, 1992)). In this kind of picture, what we suggest is that on easier complexity trials, signals coming from cortical working memory and/or hippocampal episodic memory to the striatum are low in uncertainty/noise. These signals, along a certain probability distribution, would have the capacity to tip the Q_{S1} vs Q_{S0} balance away from the striatum's local and historical valuations of these options. However, as uncertainty increases, the signals coming from these upstream memory systems would become noisier and thus less able to influence striatal action selection. The same can be suggested with respect to a putative amygdalar contribution to decision making, especially post-initial choice cognition, since the parameters we have related to this, i.e. run time and choice revision, actually tend to be strongest on easier trials in classical EdM but not in EdRR, where they are equally distributed across trial complexity levels.

If, as we suggest, this multiple memory systems interpretation of myside confirmation bias-like behavior is accurate, then what psychology refers to as myside bias may be fundamentally a latent neurobiological fact of all organisms which have evolved just such multiple memory systems; stimulated, co-opted, exapted into action in the case of humans by the specific epistemic environments of belief, knowledge, persuasion, and indoctrination we have created for ourselves. We believe that our present study has opened the way to much future research in this direction.

Bibliography

Al Abed, A. S., Sellami, A., Brayda-Bruno, L., Lamothe, V., Noguès, X., Potier, M., Bennetau-Pelissero, C., & Marighetto, A. (2016). Estradiol enhances retention but not organization of hippocampus-dependent memory in intact male mice. *Psychoneuroendocrinology*, *69*, 77–89. https://doi.org/10.1016/j.psyneuen.2016.03.014

Albayram, O., Alferink, J., Pitsch, J., Piyanova, A., Neitzert, K., Poppensieker, K., Mauer, D., Michel, K., Legler, A., Becker, A., Monory, K., Lutz, B., Zimmer, A., & Bilkei-Gorzo, A. (2011). Role of CB1 cannabinoid receptors on GABAergic neurons in brain aging. *Proceedings of the National Academy of Sciences*, *108*(27), 11256–11261. https://doi.org/10.1073/pnas.1016442108

Albayram, O., Passlick, S., Bilkei-Gorzo, A., Zimmer, A., & Steinhäuser, C. (2016). Physiological impact of CB1 receptor expression by hippocampal GABAergic interneurons. *Pflügers Archiv - European Journal of Physiology*, *468*(4), 727–737. https://doi.org/10.1007/s00424-015-1782-5

Alós-Ferrer, C., Hügelschäfer, S., & Li, J. (2016). Inertia and Decision Making. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/article/10.3389/fpsyg.2016.00169

Amsel, A. (1958). The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin*, *55*(2), 102–119. https://doi.org/10.1037/h0043125

Anderson, M. C., & Floresco, S. B. (2021). Prefrontal-hippocampal interactions supporting the extinction of emotional memories: The retrieval stopping model. *Neuropsychopharmacology*, 1–16. https://doi.org/10.1038/s41386-021-01131-1

Anderson, M. C., & Hulbert, J. C. (2021). Active Forgetting: Adaptation of Memory by Prefrontal Control. *Annual Review of Psychology*, *72*(1), 1–36. https://doi.org/10.1146/annurev-psych-072720-094140

Aultman, J. M., & Moghaddam, B. (2001). Distinct contributions of glutamate and dopamine receptors to temporal aspects of rodent working memory using a clinically relevant task. *Psychopharmacology*, *153*(3), 353–364. https://doi.org/10.1007/s002130000590

Bacon, F. (1620). The New Organon: Or True Directions Concerning the Interpretation of Nature. *The New Organon*, 134.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Ballard, I., Miller, E. M., Piantadosi, S. T., Goodman, N. D., & McClure, S. M. (2018). Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning. *Cerebral Cortex*, *28*(11), 3965–3975. https://doi.org/10.1093/cercor/bhx259

Balzan, R., Delfabbro, P., Galletly, C., & Woodward, T. (2013). Confirmation biases across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *British Journal of Clinical Psychology*, *52*(1), 53–69. https://doi.org/10.1111/bjc.12000

Bedre, R. (2021). *reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit*. Zenodo. https://doi.org/10.5281/zenodo.4422035

Bekinschtein, P., Weisstaub, N. V., Gallo, F., Renner, M., & Anderson, M. C. (2018). A retrieval-specific mechanism of adaptive forgetting in the mammalian brain. *Nature Communications*, *9*(1), 4660. https://doi.org/10.1038/s41467-018-07128-7

Bilkei-Gorzo, A., Albayram, O., Draffehn, A., Michel, K., Piyanova, A., Oppenheimer, H., Dvir-Ginzberg, M., Rácz, I., Ulas, T., Imbeault, S., Bab, I., Schultze, J. L., & Zimmer, A. (2017). A chronic low dose of Δ9-tetrahydrocannabinol (THC) restores cognitive function in old mice. *Nature Medicine*, *23*(6), 782–787. https://doi.org/10.1038/nm.4311

Chierchia, G., Soukupová, M., Kilford, E. J., Griffin, C., Leung, J. T., Blakemore, S.-J., & Palminteri, S. (2021). *Choice-confirmation bias in reinforcement learning changes with age during adolescence*. PsyArXiv. https://doi.org/10.31234/osf.io/xvzwb

Cisek, P., & Hayden, B. Y. (2022). Neuroscience needs evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200518. https://doi.org/10.1098/rstb.2020.0518

Costanzi, M., Cianfanelli, B., Santirocchi, A., Lasaponara, S., Spataro, P., Rossi-Arnaud, C., & Cestari, V. (2021). Forgetting Unwanted Memories: Active Forgetting and Implications for the Development of Psychological Disorders. *Journal of Personalized Medicine*, *11*(4), 241. https://doi.org/10.3390/jpm11040241

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific Reports*, *7*(1), 40391. https://doi.org/10.1038/srep40391

Doll, B. B., Waltz, J. A., Cockburn, J., Brown, J. K., Frank, M. J., & Gold, J. M. (2014). Reduced susceptibility to confirmation bias in schizophrenia. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 715–728. https://doi.org/10.3758/s13415-014-0250-6

Farahbakhsh, Z. Z., & Siciliano, C. A. (2021). Neurobiology of novelty seeking. *Science*, *372*(6543), 684–685. https://doi.org/10.1126/science.abi7270

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Geraerts, E., Merckelbach, H., Jelicic, M., & Habets, P. (2007). Suppression of Intrusive Thoughts and Working Memory Capacity in Repressive Coping. *The American Journal of Psychology*, *120*(2), 205. https://doi.org/10.2307/20445395

Granero, R., Fernández-Aranda, F., Valero-Solís, S., Pino-Gutiérrez, A. del, Mestre-Bach, G., Baenas, I., Contaldo, S. F., Gómez-Peña, M., Aymamí, N., Moragas, L., Vintró, C., Mena-Moreno, T., Valenciano-Mendoza, E., Mora-Maltas, B., Menchón, J. M., & Jiménez-Murcia, S. (2020). The influence of chronological age on cognitive biases and impulsivity levels in male patients with gambling disorder. *Journal of Behavioral Addictions*, *9*(2), 383–400. https://doi.org/10.1556/2006.2020.00028

Han, J., Kesner, P., Metna-Laurent, M., Duan, T., Xu, L., Georges, F., Koehl, M., Abrous, D. N., Mendizabal-Zubiaga, J., Grandes, P., Liu, Q., Bai, G., Wang, W., Xiong, L., Ren, W., Marsicano, G., & Zhang, X. (2012). Acute Cannabinoids Impair Working Memory through Astroglial CB1 Receptor Modulation of Hippocampal LTD. *Cell*, *148*(5), 1039–1050. https://doi.org/10.1016/j.cell.2012.01.037

Hu, X., Bergström, Z. M., Gagnepain, P., & Anderson, M. C. (2017). Suppressing Unwanted Memories Reduces Their Unintended Influences. *Current Directions in Psychological Science*, *26*(2), 197–206. https://doi.org/10.1177/0963721417689881

Hulbert, J. C., Henson, R. N., & Anderson, M. C. (2016). Inducing amnesia through systemic suppression. *Nature Communications*, *7*(1), 11003. https://doi.org/10.1038/ncomms11003

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Inglis, I. R., Langton, S., Forkman, B., & Lazarus, J. (2001). An information primacy model of exploratory and foraging behaviour. *Animal Behaviour*, *62*(3), 543–557. https://doi.org/10.1006/anbe.2001.1780

Jobson, D. D., Hase, Y., Clarkson, A. N., & Kalaria, R. N. (2021). The role of the medial prefrontal cortex in cognition, ageing and dementia. *Brain Communications*, *3*(3). https://doi.org/10.1093/braincomms/fcab125

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*(1), 39589. https://doi.org/10.1038/srep39589

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of Striatum in Updating Values of Chosen Actions. *Journal of Neuroscience*, *29*(47), 14701–14712. https://doi.org/10.1523/JNEUROSCI.2728-09.2009

Lara, A. H., & Wallis, J. D. (2015). The Role of Prefrontal Cortex in Working Memory: A Mini Review. *Frontiers in Systems Neuroscience*, *9*, 173. https://doi.org/10.3389/fnsys.2015.00173

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*. https://doi.org/10.1126/science.aao2998

Lustberg, D., Tillage, R. P., Bai, Y., Pruitt, M., Liles, L. C., & Weinshenker, D. (2020). Noradrenergic circuits in the forebrain control affective responses to novelty. *Psychopharmacology*, *237*(11), 3337–3355. https://doi.org/10.1007/s00213-020-05615-8

Marighetto, A., Brayda-Bruno, L., & Etchamendy, N. (2011). Studying the Impact of Aging on Memory Systems: Contribution of Two Behavioral Models in the Mouse. In M.-C. Pardon & M. W. Bondi (Eds.), *Behavioral Neurobiology of Aging* (Vol. 10, pp. 67–89). Springer Berlin Heidelberg. https://doi.org/10.1007/7854_2011_151

Markowitz, J. E., Gillis, W. F., Beron, C. C., Neufeld, S. Q., Robertson, K., Bhagat, N. D., Peterson, R. E., Peterson, E., Hyun, M., Linderman, S. W., Sabatini, B. L., & Datta, S. R. (2018). The Striatum Organizes 3D Behavior via Moment-to-Moment Action Selection. *Cell*, *174*(1), 44-58.e17. https://doi.org/10.1016/j.cell.2018.04.019

Martín-García, E., Fernández-Castillo, N., Burokas, A., Gutiérrez-Cuesta, J., Sánchez-Mora, C., Casas, M., Ribasés, M., Cormand, B., & Maldonado, R. (2015).

Frustrated expected reward induces differential transcriptional changes in the mouse brain: Microarrays and frustration. *Addiction Biology*, *20*(1), 22–37. https://doi.org/10.1111/adb.12188

McDonald, R. J., Foong, N., & Hong, N. S. (2004). Incidental information acquired by the amygdala during acquisition of a stimulus-response habit task. *Experimental Brain Research*, *159*(1), 72–83. https://doi.org/10.1007/s00221-004-1934-x

McDonald, R. J., & Hong, N. S. (2004). A dissociation of dorso-lateral striatum and amygdala function on the same stimulus–response habit task. *Neuroscience*, *124*(3), 507–513. https://doi.org/10.1016/j.neuroscience.2003.11.041

McDonald, R. J., & White, N. M. (1993). *A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. - PsycNET.* https://content.apa.org/doiLanding?doi=10.1037%2F0735-7044.107.1.3

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason.* Harvard University Press.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., Watanabe, M., Yamanaka, K., Enomoto, K., Chiken, S., Sano, H., Soma, S., Yoshida, J., Samejima, K., Ogawa, M., Kobayashi, K., Nambu, A., Isomura, Y., & Kimura, M. (2018). Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. *Neuron*, *99*(6), 1302-1314.e5. https://doi.org/10.1016/j.neuron.2018.08.002

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature*, *526*(7572), 182–185. https://doi.org/10.1038/526182a

Palminteri, S. (2021). *Choice-confirmation bias and gradual perseveration in human reinforcement learning.*

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Park, A. J., Harris, A. Z., Martyniuk, K. M., Chang, C.-Y., Abbas, A. I., Lowes, D. C., Kellendonk, C., Gogos, J. A., & Gordon, J. A. (2021). Reset of hippocampal–prefrontal circuitry facilitates learning. *Nature*, *591*(7851), 615–619. https://doi.org/10.1038/s41586-021-03272-1

Pinker, S. (2021). *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. Penguin.

Prochaska, J. O. (2008). Decision Making in the Transtheoretical Model of Behavior Change. *Medical Decision Making*, *28*(6), 845–849. https://doi.org/10.1177/0272989X08327068

Reback, J., McKinney, W., jbrockmendel, Bossche, J. V. den, Augspurger, T., Cloud, P., gfyoung, Sinhrks, Klein, A., Roeschke, M., Hawkins, S., Tratner, J., She, C., Ayd, W., Petersen, T., Garcia, M., Schendel, J., Hayden, A., MomIsBestFriend, … Mehyar, M. (2020). *pandas-dev/pandas: Pandas 1.0.3*. Zenodo. https://doi.org/10.5281/zenodo.3715232

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17*(3), 147–159. https://doi.org/10.1038/nrn.2015.30

Rollwage, M., & Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1822), 20200131. https://doi.org/10.1098/rstb.2020.0131

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), 2634. https://doi.org/10.1038/s41467-020-16278-6

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science*, *310*(5752), 1337–1340. https://doi.org/10.1126/science.1115270

Shoji, H., Hagihara, H., Takao, K., Hattori, S., & Miyakawa, T. (2012). T-maze Forced Alternation and Left-right Discrimination Tasks for Assessing Working and Reference Memory in Mice. *JoVE (Journal of Visualized Experiments)*, *60*, e3300. https://doi.org/10.3791/3300

Stanovich, K. E. (2021). *The Bias That Divides Us: The Science and Politics of Myside Thinking*. MIT Press.

Stanovich, K., & West, R. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning - THINK REASONING*, *13*, 225–247. https://doi.org/10.1080/13546780600780796

Stanovich, K., West, R., & Toplak, M. (2013). Myside Bias, Rational Thinking, and Intelligence. *Current Directions in Psychological Science*, *22*, 259–264. https://doi.org/10.1177/0963721413480174

Stevens, C., Lacroix, C., Mariani, Y., Marsicano, G., & Marighetto, A. (2022a). Indoctrination as active inhibition of spontaneous exploration: Introduction of a novel mouse model. (In preparation)

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). A Bradford Book.

Takeuchi, T., Duszkiewicz, A. J., Sonneborn, A., Spooner, P. A., Yamasaki, M., Watanabe, M., Smith, C. C., Fernández, G., Deisseroth, K., Greene, R. W., & Morris, R. G. M. (2016). Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature*, *537*(7620), 357–362. https://doi.org/10.1038/nature19325

Taylor-Rolph Co. Ltd. (1938). *What is bias?* Taylor-Rolph Co. Ltd.

Terra, H., Bruinsma, B., de Kloet, S. F., van der Roest, M., Pattij, T., & Mansvelder, H. D. (2020). Prefrontal Cortical Projection Neurons Targeting Dorsomedial Striatum Control Behavioral Inhibition. *Current Biology*, *30*(21), 4188-4200.e5. https://doi.org/10.1016/j.cub.2020.08.031

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wason, P. C. (1960). On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Wason, P. C. (1966). *New Horizons in Psychology*. Penguin Books.

Wason, P. C. (1968). Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292. https://doi.org/10.1007/BF00992698

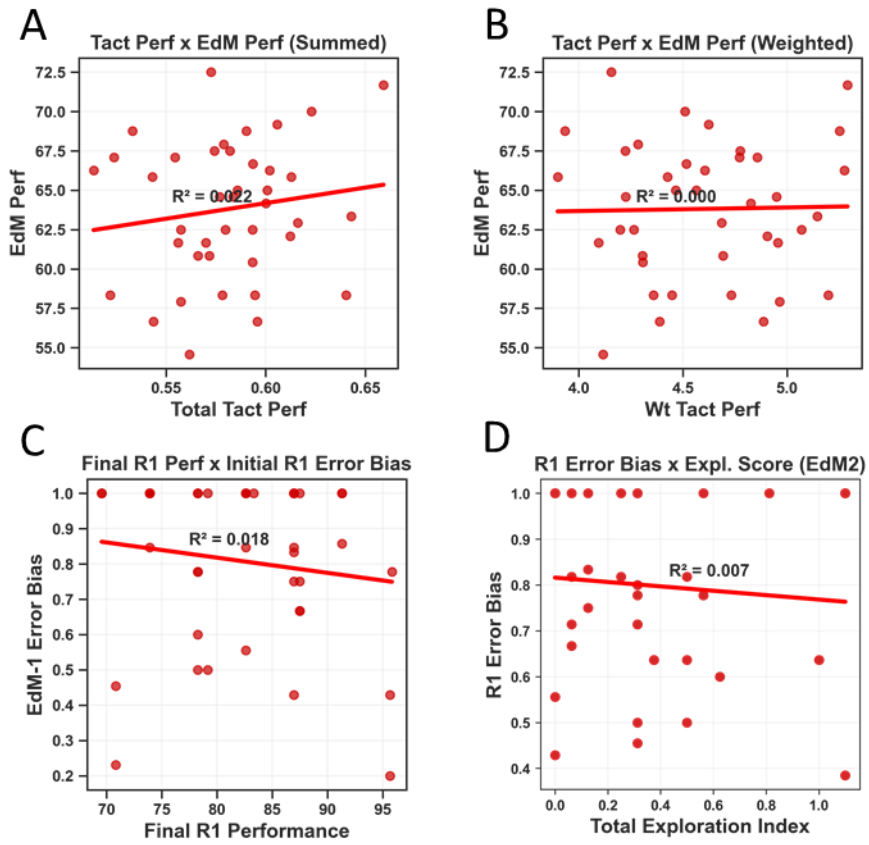
White, N. M., & McDonald, R. J. (2002). Multiple parallel memory systems in the brain of the rat. *Neurobiology of Learning and Memory*, *77*(2), 125–184. https://doi.org/10.1006/nlme.2001.4008

Wilson, C. G., Nusbaum, A. T., Whitney, P., & Hinson, J. M. (2018). Age-differences in cognitive flexibility when overcoming a preexisting bias through feedback.

Journal of Clinical and Experimental Neuropsychology, *40*(6), 586–594. https://doi.org/10.1080/13803395.2017.1398311

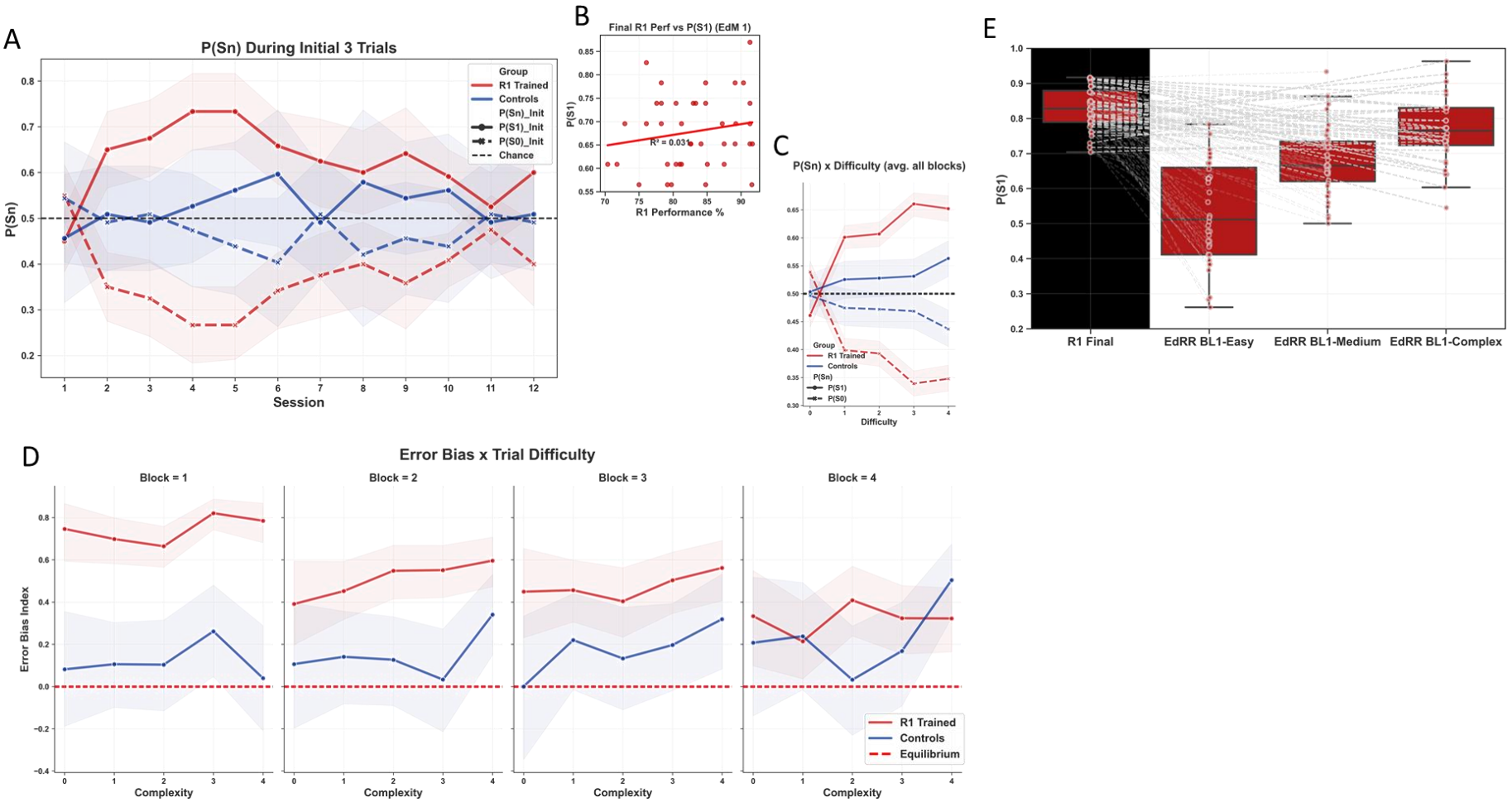
Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, *17*(6), e1009070. https://doi.org/10.1371/journal.pcbi.1009070

Supplementary Figures S.1-3:



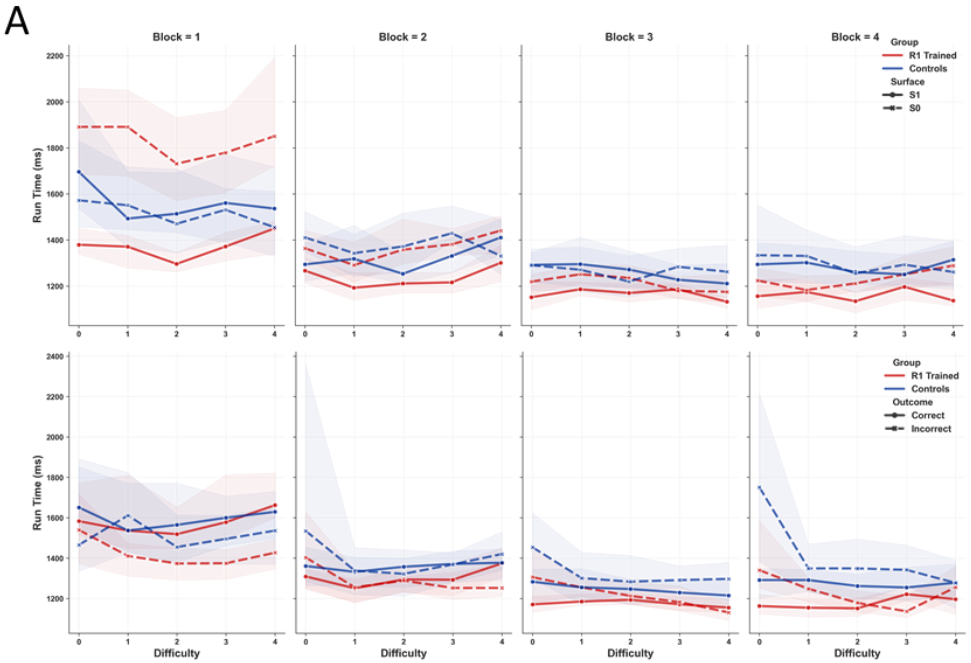
Supplementary figure S. 5 - Absence of correlations hints towards interaction of multiple cognitive systems.

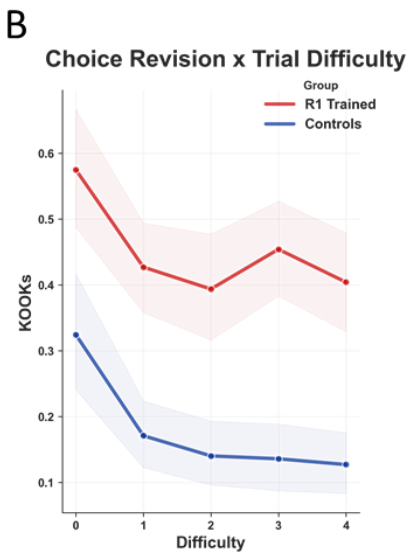
Several linear regression analyses were calculated from individual R1 trained population parameters in order to look for correlations between behaviors which intuitively could have been correlated. (**A**) No correlation was found between summed tactile discrimination (R1) performance in the training phase and summed EdM performance during EdRR (a *negative* correlation might have been expected). (**B**) We then weighted the R1 performance, discounting earlier compared to later performances, in case earlier low R1 performances were occluding a relationship, but still no correlation between these two behaviors. (**C**) Specifically between R1 performance in the final R1 session only and within-error R1 bias during the first EdRR session only (two sessions which were separated by 24 hours), we again found no correlation. Notably, both relatively weak and strong R1 performers were just as likely to display a 100% S1 error bias during EdRR 1. (**D**) Finally, having observed that novelty per se had a major impact on behavior, especially during EdRR 1, we instead compared individual EI values in EdRR 2 with levels of within-error R1 bias also in EdRR 2, but again no correlation was seen.

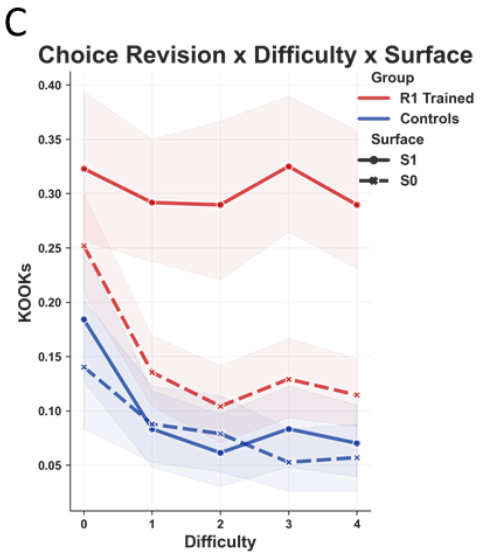

A
P(Sn) During Initial 3 Trials
Group
R1 Trained
Controls
P(Sn)_Init
P(S1)_Init
P(S0)_Init
Chance
P(Sn)
Session
B
Final R1 Perf vs P(S1) (EdM 1)
P(S1)
R² = 0.031
R1 Performance %
C
P(Sn) x Difficulty (avg. all blocks)
Group
R1 Trained
Controls
P(Sn)
P(S1)
P(S0)
Difficulty
D
Error Bias x Trial Difficulty
Block = 1
Block = 2
Block = 3
Block = 4
Error Bias Index
Complexity
R1 Trained
Controls
Equilibrium
E
P(S1)
R1 Final
EdRR BL1-Easy
EdRR BL1-Medium
EdRR BL1-Complex

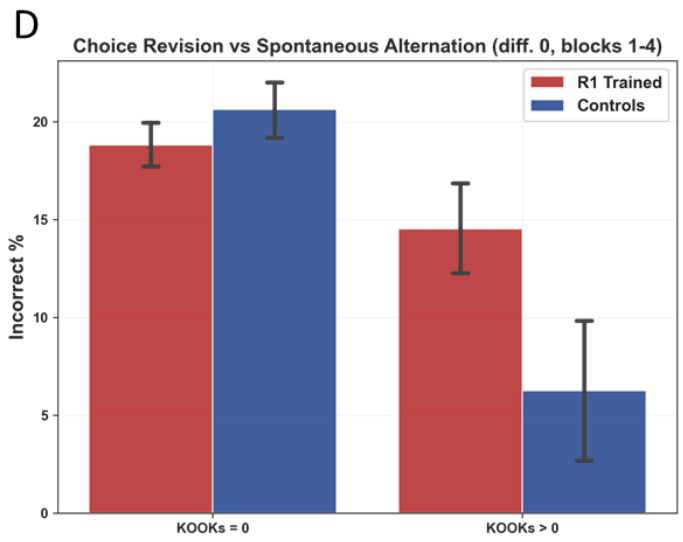
Supplementary figure S. 6 – S1 bias clear and robust during EdRR, as a relative difference within errors and as a function of trial complexity in overall P(S1).

All error bands represent 95% confidence intervals, vertical spaces between bands indicate statistical significance (detailed statistical analyses in main text). (**A**) Isolated P(Sn) values for initial 3 trials which are omitted by analyses by trial complexity. Once again, the impact of novelty on initial exploratory behavior in R1 trained animals during EdRR 1 is very clear. Following EdRR 1, P(S1) is significantly higher than P(S0) in almost all sessions in the R1 trained population. In controls, we observed more fluctuation, but still with a slight tendency for P(S1) to be higher. Recall that in these initial 3 trials, i.e. the first trial on each pair, only S1 was rewarded, thus P(Rw|S1) = 1 and P(Rw|S0) = 0 for these trials. (**B**) Linear regression analysis found no correlation between final R1 performance and overall P(S1) in EdRR 1. (**C**) P(Sn) as a function of trial complexity averaged across all sessions/blocks. In this representation, it is clear that P(S1) is, on average, slightly higher in controls on all trial complexities above level 0, but this reaches significance only on level 4 trials. (**D**) As we had done for the overall within-error R1 bias index, we also calculated the relative difference in S1 versus S0 errors as a function of trial complexity. At this level, there was no dependence of within-error R1 bias on trial complexity, only a slight dependence on time/repeated training which mirrored the trend of the overall within-error R1 bias to decrease over time. However, in controls, we can again see that, in blocks 2-4 especially, this population was significantly more likely to make errors on S1 than on S0 arms on complexity level 4 trials. (**E**) Having seen from figure 4 that P(Sn) in R1 trained animals seemed to mark out 3 distinct groupings of complexity (level 0, easy; levels 1-2, medium; levels 3-4, complex), we looked at P(S1) in the final R1 session (which was equal, by definition of the task, to R1 performance) and compared it to P(S1) values across the first block of 3 EdRR sessions grouped according to these three levels. This provides a clear visual representation of just how close R1 trained animals were in early EdRR sessions to performing on complex trials almost identically to how they had performed in the R1 environment where they had never experienced any R1 disconfirmations.


A
Block = 1
Block = 2
Block = 3
Block = 4
Group
R1 Trained
Controls
Surface
S1
S0
Outcome
Correct
Incorrect
Run Time (ms)
Difficulty


B
Choice Revision x Trial Difficulty
Group
R1 Trained
Controls
KOOKs
Difficulty


C
Choice Revision x Difficulty x Surface
Group
R1 Trained
Controls
Surface
S1
S0
KOOKs
Difficulty


D
Choice Revision vs Spontaneous Alternation (diff. 0, blocks 1-4)
R1 Trained
Controls
Incorrect %
KOOKs = 0
KOOKs > 0

Supplementary figure S. 7 – R1 bias disturbs development of normal EdM phenotype in deliberative behaviors.

All error bands represent 95% confidence intervals. (**A**) top; median run time by surface by difficulty by block. (**A**) bottom; median run time by outcome by difficulty by block. As shown above, R1 trained animals maintain persistently lower run times on S1 in all blocks and at all trial complexity levels. As a function of outcome, we can see that controls in EdRR more quickly develop a stable run time phenotype across all trial complexities (i.e. higher run times on incorrect outcome trials; see supplementary figures A.3 + A.6) compared to R1 trained animals, which show a lag in this respect on more complex trials especially. This could, precisely, be caused by a sustained "over-confidence" effect on S1 choices, independently of outcome. (**B**) Mean total choice revision by difficulty. Averaged across all blocks, R1 trained animals displayed significantly more choice revision than controls at all trial complexities. However, both groups engaged in significantly more choice revision behavior on level 0 trials. (**C**) Mean total choice revision by difficulty by surface. Dividing mean choice revision values according to surface further reveals that choice revision in R1 trained animals, but not controls, that terminated on S1 surfaces was equally high on all trial complexity levels, or rather was just as high on level 1-4 trials as it was on level 0 trials. This indicates that striatal R1 cognitive content has a greater capacity to remain active and potentially intrusive at any trial complexity level, in contrast to cortico-hippocampal EdM content which decreases in activation (putatively as an effect of adaptive forgetting), thus leaving fewer EdM elements capable of provoking choice revision, as a function of complexity level. (**D**) In level 0 trials, almost all choice revision in controls was rectifying whereas R1 trained animals were comparatively likely to make an error with as without choice revision. 88% of these errors *with* choice revision, averaged across all blocks, were from S1 final choices (see main text).

Appendix / Supplementary material

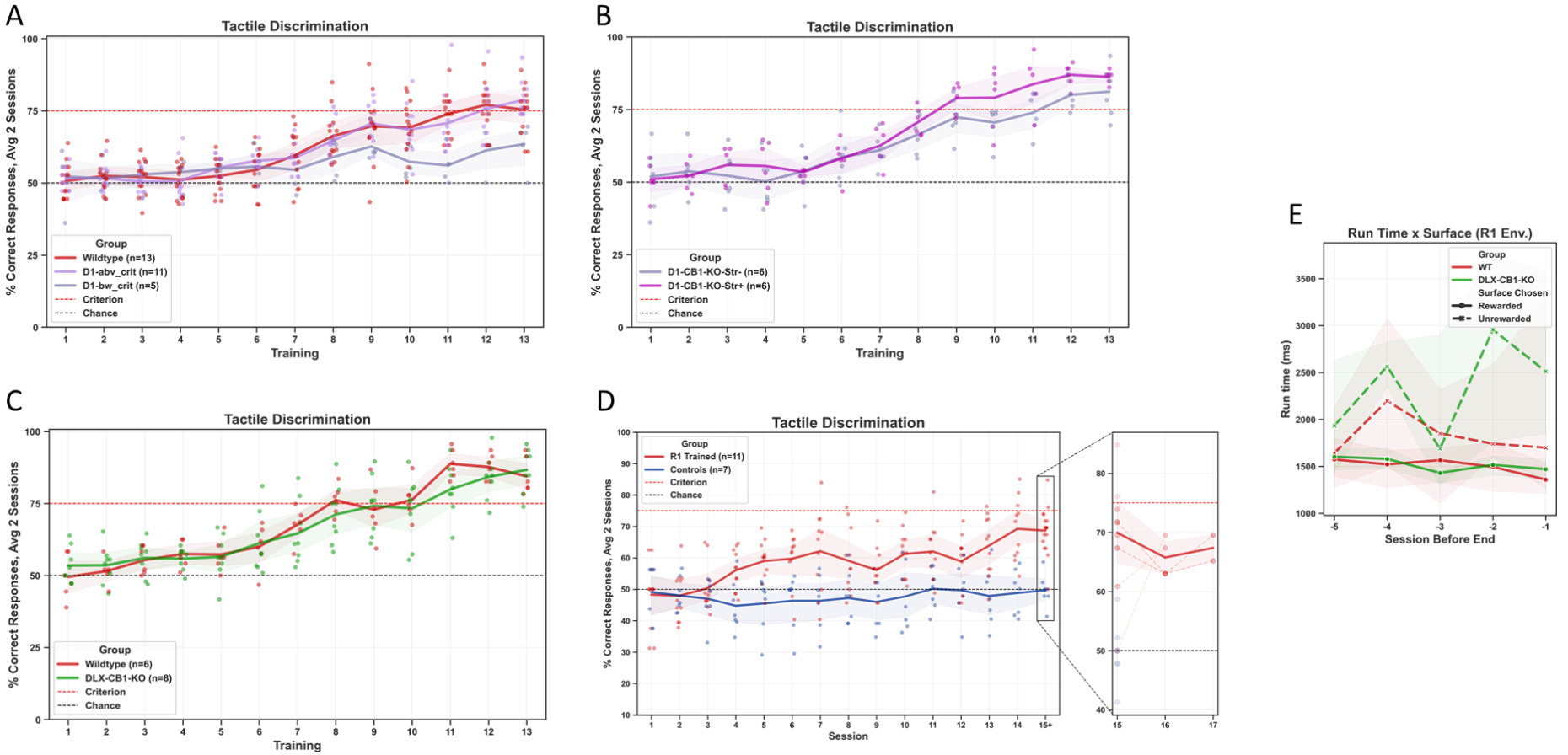
Preliminary experimental interventions.

Having characterized our novel protocol and demonstrated its face validity as a model of both rule revision and 'myside' confirmation bias in an everyday-like memory-based cognitive task, we next employed both genetic and ageing approaches to test its practical potential as a tool for dissecting and better understanding the underlying cognitive and neurobiological mechanisms. In this secondary results section, we briefly summarize the data collected from a selection of exploratory experiments, primarily with a view to demonstrating the model's potential to both inspire novel research questions and carry forward established domains of research, such as the impacts of ageing on cognition. These results are referenced where appropriate in the discussion of the main paper.

A.1 Genetic approaches.

A.1.1 Independence of R1 acquisition and expression demonstrated via manipulation of striatal function.

As seen and discussed in detail in Stevens et al., 2022a, D_1-CB_1-KO mice (a mouse breed in which CB_1 has been conditionally deleted from dopamine type-I receptor (D_1) expressing neurons of the forebrain) were impaired in R1 expression but not acquisition compared to wildtype littermates. Within the striatum, D_1 are expressed on inhibitory GABAergic medium spiny neurons of the direct pathway. CB_1 expressed on the pre-synaptic element of these neurons exert a retrograde inhibitory modulatory effect, thus producing a net effect of inhibiting an inhibitory signal. Deletion of CB_1 receptors from these neurons has been shown to potentiate their net inhibitory signal (Soria-Gomez et al. 2021). Thus, our hypothesis for the R1 phase had been that continued selection of the spontaneous exploratory response would be potentiated in D_1-CB_1-KO animals, rendering this cognitive strategy more resistant to the active inhibition necessary to allow the R1 response strategy to be properly "chunked" (i.e. consolidated from multi-component actions into a singular, direct pathway selectable one, such as riding a bicycle or touch-typing) and thereby expressed in a sustained manner also via the direct pathway (Graybiel 1998; Jin, Tecuapetla, and Costa 2014).

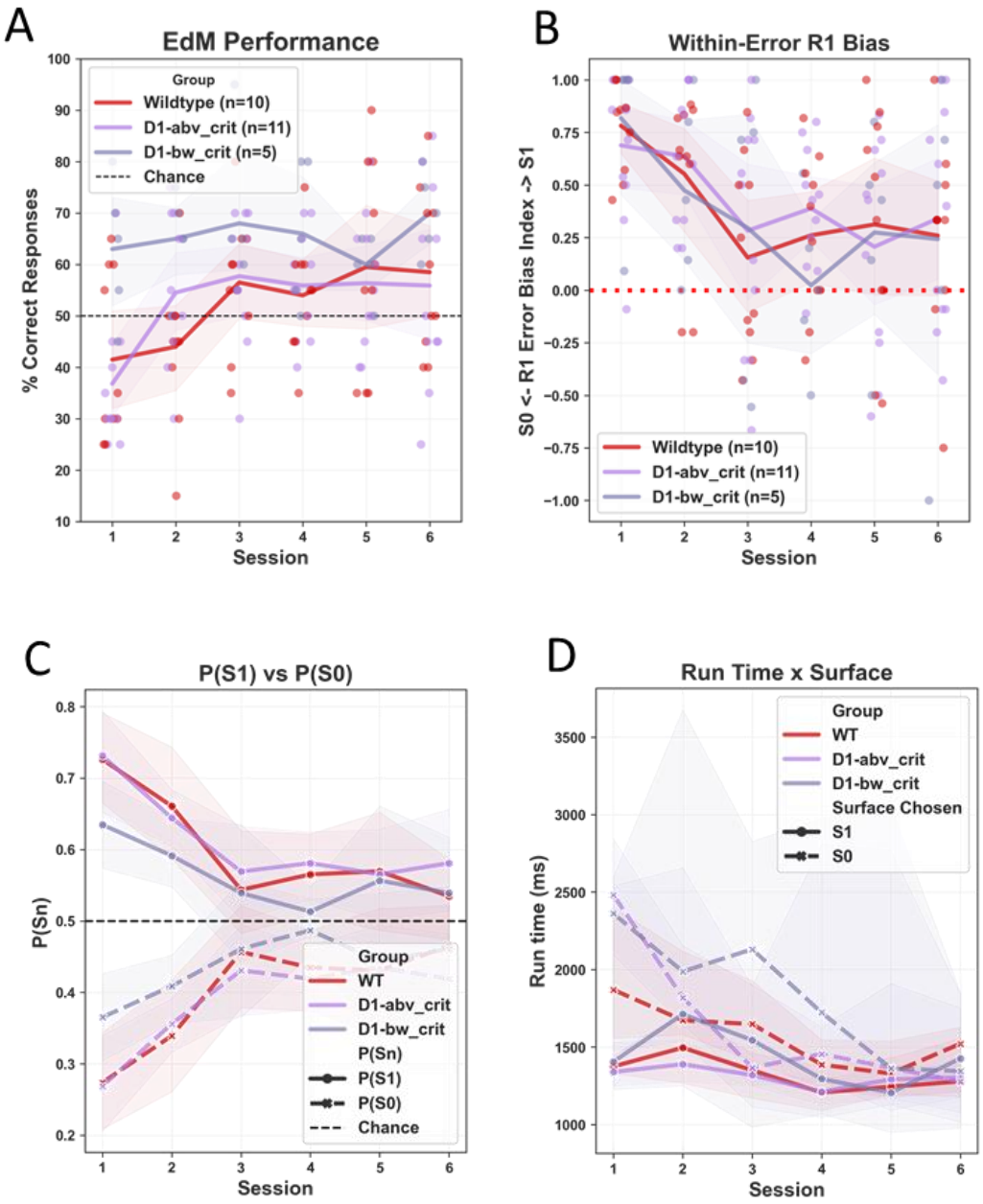

A
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Training
Group
Wildtype (n=13)
D1-abv_crit (n=11)
D1-bw_crit (n=5)
Criterion
Chance
B
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Training
Group
D1-CB1-KO-Str- (n=6)
D1-CB1-KO-Str+ (n=6)
Criterion
Chance
C
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Training
Group
Wildtype (n=6)
DLX-CB1-KO (n=8)
Criterion
Chance
D
Tactile Discrimination
% Correct Responses, Avg 2 Sessions
Session
Group
R1 Trained (n=11)
Controls (n=7)
Criterion
Chance
E
Run Time x Surface (R1 Env.)
Run time (ms)
Session Before End
Group
WT
DLX-CB1-KO
Surface Chosen
Rewarded
Unrewarded

Supplementary figure A. 1 – R1 expression performances of transgenic and aged cohorts during R1 training phase.

All error bands represent 95% confidence intervals. (**A**) Wildtype mice (red, n=13) plus D_1-CB_1-KO subdivided into two populations: D_1-CB_1-KO_{abv_crit} (, n=11), who reached R1 criterion performance of 75%, and; D_1-CB_1-KO_{bw_crit} (blue-gray, n=5), who did not. (**B**) R1 performance was restored and even reinforced in D_1-CB_1-$KO_{Str+/+}$ (magenta, n=6) who expressed and sustained R1 more rapidly and more robustly than D_1-CB_1-$KO_{Str-/-}$ mice (blue-gray, n=6). (**C**) No significant differences in R1 expression were observed between wildtype (red, n=6) and Dlx-CB_1-KO mice (green, n=8). All individuals from both groups reached criterion performance. (**D**) Aged R1 trained mice (red, n=11) required extensive training compared to young adult mice to reach R1 criterion performance. Even after 17 sessions of training, 3 of the aged individuals still showed no signs of reaching it and were thus excluded and not brought to the EdRR phase. (**E**) The only phenotype observed in Dlx-CB_1-KO mice during R1 training was their slightly but significantly longer (when averaged across the 5 final pseudo-random R1 sessions) run times on S0 choices.

R1 acquisition versus expression during EdRR: Our direct pathway potentiation hypothesis implied what we might call a threshold effect; the R1 strategy would meet more resistance in becoming "chunked", but if and when this happened, we should then no longer observe a phenotype related to its expression, since this expression would henceforth also be controlled via the direct pathway. To explore this idea, we divided the D_1-CB_1-KO mice into two groups at the end of R1 training; those that had reached criterion level R1 performance and those that had not. Those who had reached criterion we labelled D_1-CB_1-KO_{abv_crit} (n=11, including 3 borderline cases whose highest 2-session average R1 performance was 73.9% rather than 75%, but who had performed above 80% in at least one session), and those who had not reached criterion we labelled D_1-CB_1-KO_{bw_crit} (n=5) (see supplementary figure A.1a). As reported in Stevens et al., 2022a, we observed no difference in measures of R1 acquisition between these two groups, despite the significant difference in expression. At this point, we posited that, if our original hypothesis were accurate (i.e. that D_1-CB_1-KO_{bw_crit} animals failed to reach R1 criterion due to a failure to sufficiently inhibit spontaneous exploration), then in the EdRR task, which leans on the spontaneous exploration behavior of spatial alternation, we should observe stronger performances in below criterion animals compared to above criterion animals, whether D_1-CB_1-KO or wildtype. (Recall that in a pilot study, we had observed no memory related or other cognitive phenotype differences in D_1-CB_1-KO mice compared to their wildtype littermates in the classical EdM protocol, see supplementary figure A.3a-d.)

As predicted, the D_1-CB_1-KO_{bw_crit} group performed significantly better in EdRR than both above R1 criterion wildtype and D_1-CB_1-KO_{abv_crit} animals, not only at first, but reliably over the course of 6 sessions of EdRR. In contrast, there was no significant difference in EdM performance between the D_1-CB_1-KO_{abv_crit} and wildtype groups (supplementary figure A.2a; one-way ANOVAs with pairwise Tukey HSD post-hoc; D_1-CB_1-KO_{bw_crit} vs D_1-CB_1-KO_{abv_crit}, $F(1, 94) = 16.1$, $p = 0.001$; D_1-CB_1-KO_{bw_crit} vs WT, $F(1, 88) = 15.4$, $p = 0.001$; D_1-CB_1-KO_{abv_crit} vs WT, $F(1, 124) = 0.04$, $p = 0.85$). This is strong evidence in favor of our hypothesis that the cognitive mechanisms which had prevented D_1-CB_1-KO_{bw_crit} from reaching R1 criterion were indeed related to the strength of expression of the spontaneous exploratory drive which the classical EdM task was precisely designed to mobilize (Al Abed et al. 2016). However, despite this stronger EdM performance, the D_1-CB_1-KO_{bw_crit} group were equally biased towards S1



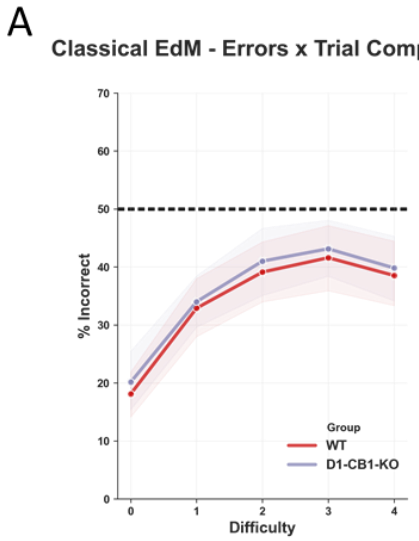
Supplementary figure A. 2 – Below R1 criterion D_1-CB_1-KO mice display significantly stronger EdM performance, but still biased.

All error bands represent 95% confidence intervals. (**A**) EdM performance during EdRR. Both wildtype (n=10) and D_1-CB_1-KO_{abv_crit} animals (n=11) displayed characteristically poor EdM performances during initial EdRR sessions. In contrast, D_1-CB_1-KO_{bw_crit} animals scored above chance level starting from EdRR 1 and reliably performed better than the other two groups across all EdRR sessions. (**B**) Within-error R1 bias results, however, revealed that, despite there being fewer of them in total, on the errors they did commit, D_1-CB_1-KO_{bw_crit} were just as biased towards S1 as their above criterion littermates, wildtype and KO. (**C**) P(Sn) values revealed that all three groups, despite different starting values, quickly dropped to and maintained similarly higher P(S1) compared to P(S0) values. (**D**) In terms of run time according to surface also, D_1-CB_1-KO_{bw_crit} animals displayed at least equal sensitivity, with comparably larger run times on S0 compared to S1 surfaces observed in all three groups.

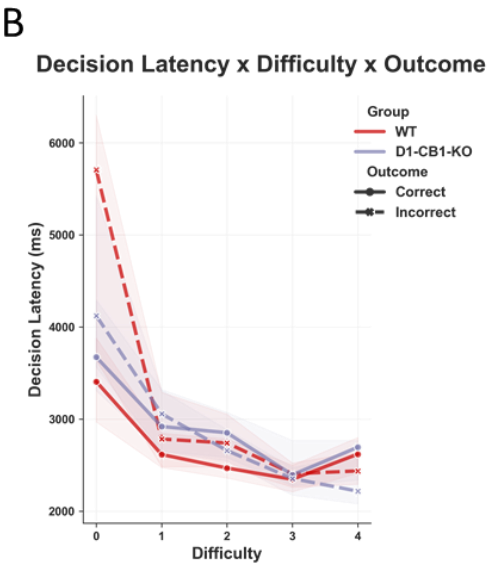
in their errors (supplementary figure A.2b) and also, albeit slightly less than the D_1-CB_1-KO_{abv_crit} and wildtype groups, in their overall P(S1) behavior (supplementary figure A.2c). Moreover, in terms of sensorial environmental feedback, when we looked at run time by surface chosen in the EdRR phase, we saw that the D_1-CB_1-KO_{bw_crit} group displayed the same characteristically significant higher run times on S0 compared to S1 (supplementary figure A.2d), indicating that the R1 bias we were observing may itself be best understood as a striatally-based matter of expression rather than a cortical question of cognitive content.

Re-expression of CB1 receptors in D1 expressing neurons of the striatum rescues capacity for exploitation of an exploration-antagonistic stimulus-response rule: In order to verify the deficit in overall R1 performance we had observed in the D_1-CB_1-KO population was indeed related to CB_1-mediated mechanisms in the direct pathway, we used a viral approach to re-express CB_1 receptors locally in D_1 expressing neurons of the striatum (D_1-CB_1-KO-$Str_{+/+}$, n = 6). We hypothesized that this re-expression would rescue expected levels of R1 exploitation compared to D_1-CB_1-KO animals injected with an empty vector virus also in the striatum (D_1-CB_1-KO-$Str_{-/-}$, n = 6; see Materials & Methods). Verifying this, in the R1 phase, we observed that D_1-CB_1-KO-$Str_{+/+}$ mice did perform significantly better in R1 than D_1-CB_1-KO-$Str_{-/-}$ mice, strongly expressing R1 both earlier and more robustly (less session-by-session variance) than D_1-CB_1-KO-$Str_{-/-}$ mice (supplementary figure A.1b), and even comparatively stronger than wildtype animals from other iterations of the experiment. Since viral re-expression of CB_1 using the CAG-promoter method generally gives rise to an over-expression relative to wildtype levels (Hitoshi, Ken-ichi, and Jun-ichi 1991), we suggested increased inhibitory modulatory control in the direct pathway, due to local CB_1 over-expression, as a putative explanation for this group's improved capacity to inhibit spontaneous exploratory behavior, thereby allowing for earlier and stronger R1 expression, via the mechanisms detailed above.

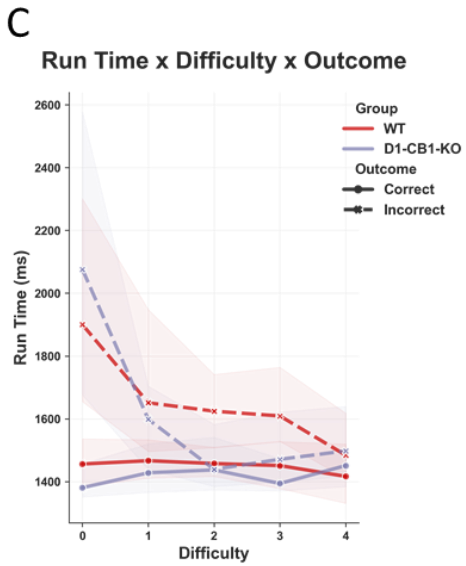
A


Classical EdM - Errors x Trial Complexity
% Incorrect
Difficulty
Group
WT
D1-CB1-KO

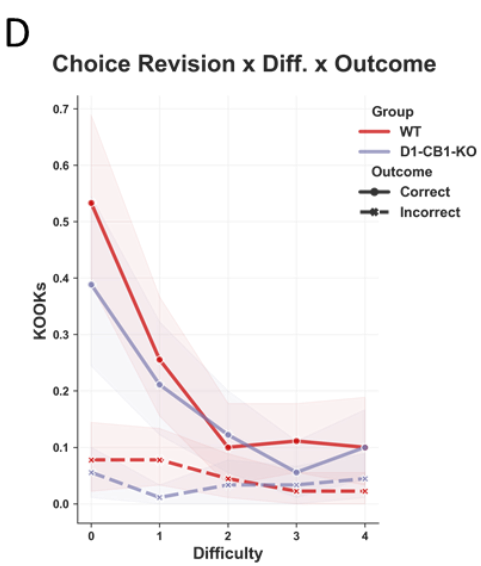
B


Decision Latency x Difficulty x Outcome
Decision Latency (ms)
Difficulty
Group
WT
D1-CB1-KO
Outcome
Correct
Incorrect

C

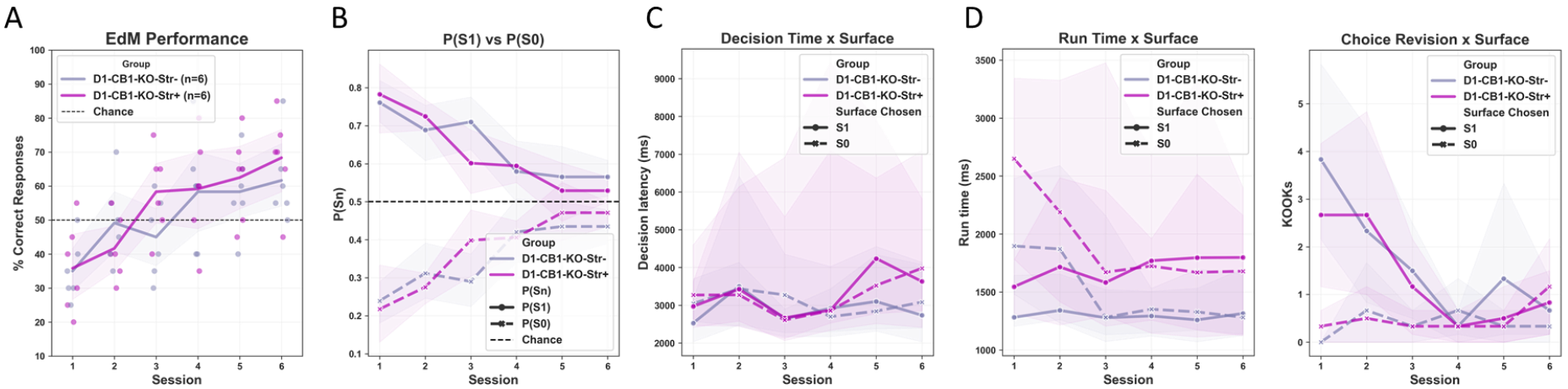

Run Time x Difficulty x Outcome
Run Time (ms)
Difficulty
Group
WT
D1-CB1-KO
Outcome
Correct
Incorrect

D


Choice Revision x Diff. x Outcome
KOOKs
Difficulty
Group
WT
D1-CB1-KO
Outcome
Correct
Incorrect

Supplementary figure A. 3 - D_1-CB_1-KO display no specific phenotype in classical EdM task.
Before characterizing how different transgenic mouse lines would respond in the EdRR task, we first ran preliminary tests to measure their performance in the classical EdM task. All error bands represent 95% confidence intervals. We also measured more detailed cognitive behaviors like decision latency and, thanks to analyses we developed for the EdRR task, also run time and choice revision. (**A**) % Incorrect responses according to difficulty. Wildtype (red, n=6) and D_1-CB_1-KO (lavender, n=6) animals displayed almost identical and characteristic EdM performances, committing least errors on level 0, and globally more errors as trial complexity increased. (**B**) Decision latency by trial complexity by outcome. Again, both groups displayed a characteristic profile of decision latencies decreasing as a function of trial complexity. We believe this highly replicable phenotype may be the result of having more detailed active cognitive content to process on easier trials compared to more complex one where, precisely, relative cognitive content has already significantly faded, as reflected in performance, and perhaps here also. Interestingly, especially on easier trials, decision latencies on trials were the outcome was an error tended to be slightly longer than when the outcome was correct. (**C**) Run time by trial complexity by outcome. This is the first time post-decision run time in the classical EdM task has been characterized. What we see is that run time when the outcome is correct is stable across all trial complexity levels but is higher on easier trials when the outcome is incorrect, decreasing then as a function of trial complexity. This indicates a higher level of hesitancy, a lower level of post-choice confidence when animals are travelling down what will be an unrewarded arm. What cognitive processes lead to this phenotype, whether it reflects "suspicion" of error or "intentional" EdM transgression, these are open questions. (**D**) Choice revision by trial complexity by outcome. Similarly, mice were more likely to revise their choice on easier trials and, overall, but especially on easier trials, choice revision was rectifying significantly more often than error-inducing. Indeed, choice revision in these two groups, as we can see, almost never resulted in error.

When moved to the EdRR environment, however, no significant differences were observed between the D_1-CB_1-KO-$Str_{-/-}$ and D_1-CB_1-KO-$Str_{+/+}$ groups in EdM performance (supplementary figure A.4a), in P(Sn) values (supplementary figure A.4b), in decision latency (supplementary figure A.4c), or in choice revision (supplementary figure A.4d). The D_1-CB_1-KO-$Str_{+/+}$ group did, however, display higher overall run times than D_1-CB_1-KO-$Str_{-/-}$ and also higher run times specifically on S0 choices (supplementary figure A.4e; one-way ANOVAs with pairwise Tukey HSD post-hoc, between 'Group', $F(1, 142) = 19.6$, $p = 0.001$; within 'Surface' [S0] between 'Group', $F(1, 70) = 10$, $p = 0.003$). This indicated that the D_1-CB_1-KO-$Str_{+/+}$ group retained higher cognitive sensitivity and responsivity (putatively via increased CB_1 modulatory activity) to tactile-based reward location probabilities grounded in striatal S-R mechanisms, even though this did not have a manifest effect on their EdM choice behavior. This may also constitute further, retrospective evidence that what allowed these animals to attain and maintain higher R1 expression in the first phase was indeed a case of ongoing active mechanisms of inhibition, rather than a case of simply reaching a "state" of stronger R1 responding.

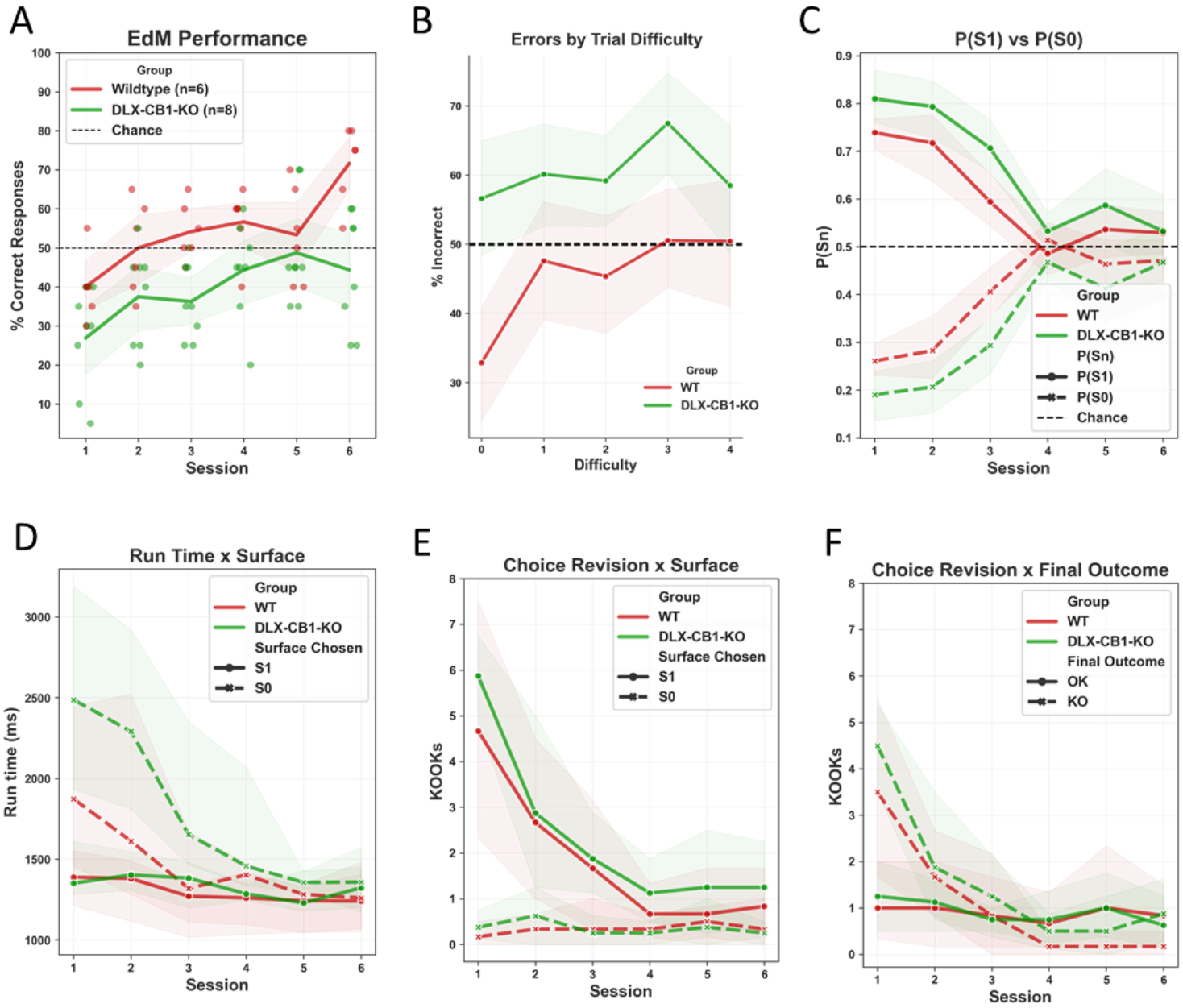

A
EdM Performance
% Correct Responses
Session
Group
D1-CB1-KO-Str- (n=6)
D1-CB1-KO-Str+ (n=6)
Chance
B
P(S1) vs P(S0)
P(Sn)
Session
Group
D1-CB1-KO-Str-
D1-CB1-KO-Str+
P(Sn)
P(S1)
P(S0)
Chance
C
Decision Time x Surface
Decision latency (ms)
Session
Group
D1-CB1-KO-Str-
D1-CB1-KO-Str+
Surface Chosen
S1
S0
D
Run Time x Surface
Run time (ms)
Session
Group
D1-CB1-KO-Str-
D1-CB1-KO-Str+
Surface Chosen
S1
S0
Choice Revision x Surface
KOOKs
Session
Group
D1-CB1-KO-Str-
D1-CB1-KO-Str+
Surface Chosen
S1
S0

Supplementary figure A. 4 - Re-expression of CB1 locally in striatum of D_1-CB_1-KO mice had no impact on EdRR performance.

During R1 training, we had observed that D_1-CB_1-$KO_{Str+/+}$ mice expressed R1 more rapidly and more robustly than their D_1-CB_1-$KO_{Str-/-}$ littermates. However, when we moved them to the EdRR environment, they displayed neither worse EdM performance nor higher bias than their littermates. All error bands represent 95% confidence intervals. (**A**) % Correct EdM responses. D_1-CB_1-$KO_{Str+/+}$ performed equally poorly initially as their D_1-CB_1-$KO_{Str-/-}$ littermates and improved their EdM performance at a similar rate. (**B**) P(Sn). In terms of surface choice probability, both groups showed equally significant preference for choosing S1 over S0 and displayed comparable decrease in this preference with repeated EdRR training. (**C**) Median decision latencies by group and by surface. Both groups displayed comparable decision latencies with no significant difference according to surface chosen. Values were, however, highly variable, especially in D_1-CB_1-$KO_{Str+/+}$ animals. (**D**) Median run times by group and by surface. D_1-CB_1-$KO_{Str+/+}$ displayed consistently higher median values in run time on both S1 and S0 surfaces compared to D_1-CB_1-$KO_{Str-/-}$ littermates, once again with very high variability. (**E**) Mean total choice revision by group and by surface. Both groups displayed significant S1 favoring choice revision behavior, which decreased at a similar rate in both populations with repeated EdRR training.

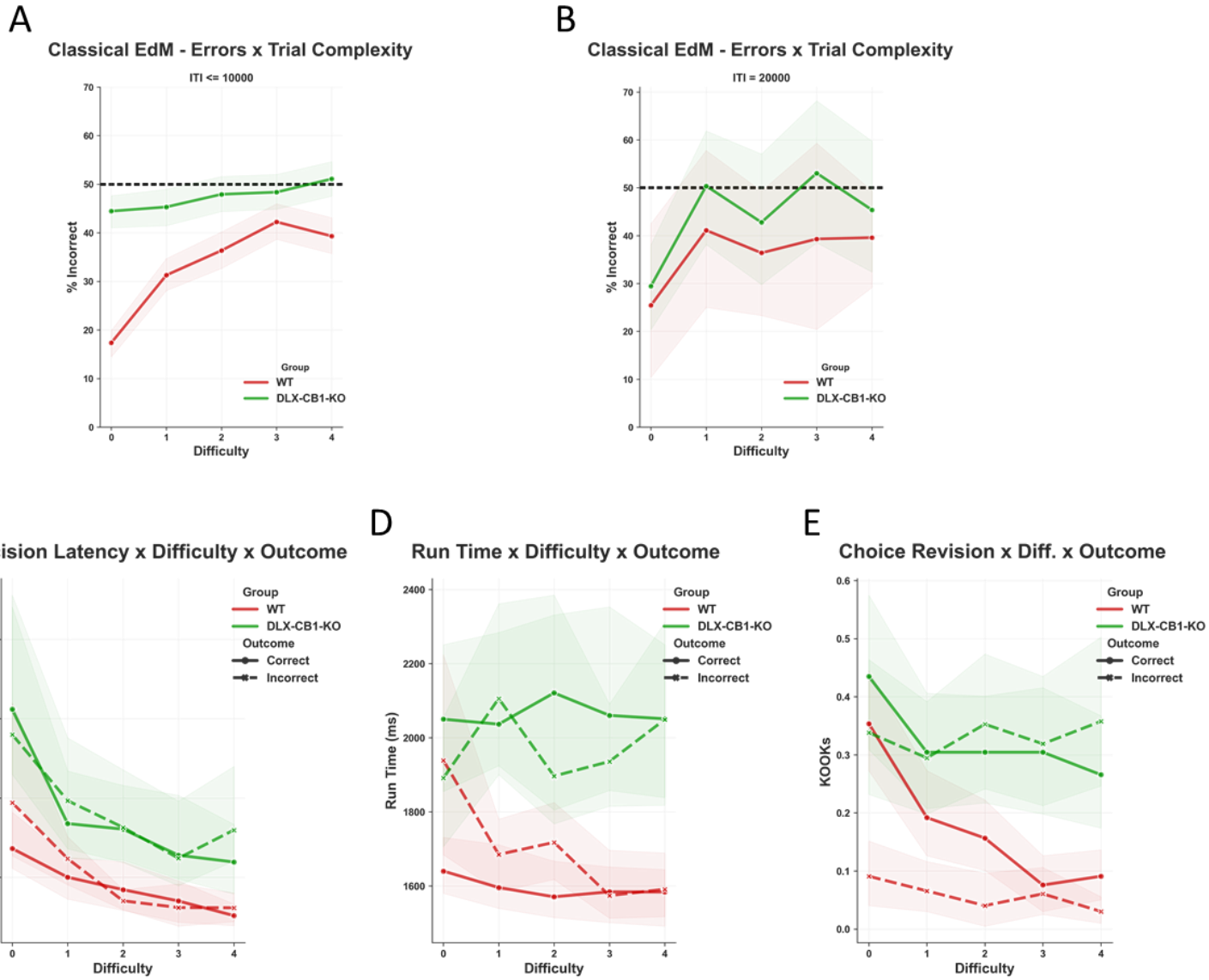
A.1.2 Inhibition of R1 interference independent of successful EdM performance.

In order to begin disentangling hippocampal from striatal contributions to R1 rule revision in the EdRR environment, we turned to the Dlx-CB_1-KO transgenic mouse line, in which CB_1 receptors are conditionally deleted from all GABAergic neurons of the forebrain. In preliminary work, we had validated the hypothesis that Dlx-CB_1-KO mice would be severely impaired in the hippocampus-dependent classical EdM task (supplementary figure A.6a-b). This hypothesis had been based in part on work showing that CB_1 expressed on GABAergic neurons of the hippocampus are necessary for the formation of second-degree indirect associations but not first-degree direct associations (Busquets-Garcia et al. 2018). Hypothesizing that this impairment could be caused by a dysfunction in the specifically organizational and active inhibition/adaptive forgetting dimensions of cortico-hippocampal memory (two key cognitive elements of the classical EdM paradigm), we had therefore predicted that Dlx-CB_1-KO would be incapable of successfully performing the EdM task, albeit in a task complexity-dependent manner. Specifically, we predicted they would achieve better than chance performance on level 0 trials, since we had also verified, confirming the literature (Albayram et al. 2016), that this mouse line had no deficit in basic Y-maze spontaneous alternation (data not shown). Moreover, previous work from our lab had shown that animals lesioned in the CA1 region of the hippocampus had relatively normal performances on level 0 complexity trials of the EdM task but did not reach better than chance performance on any of the more complex levels (Marighetto, Brayda-Bruno, and Etchamendy 2011). Looking at supplementary figure A.6a, the deficit in performance in Dlx-CB_1-KO mice compared to their wildtype littermates is clear and striking. Beyond this between group difference, we can see a clear effect of 'Difficulty' in the wildtype group (one-way ANOVA, $F(4, 985) = 30.1$, $p < 0.0001$; pairwise Tukey HSD post-hoc tests revealed statistically significant differences in 6 out of 10 levels of comparison). Looking closer only at the Dlx-CB_1-KO results, level 0 was slightly but significantly above chance performance (t-test with Welch correction and unbiased Cohen effect size between complexity level 0 performance and chance level of 50%, $t(206) = 3.3$, $p = 0.001$, $d = 0.23$) but when we ran an ANOVA to check the statistical significance of the slight effect of Difficulty visible on the graph, this trend did not achieve significance at any level of


A
EdM Performance
Group
Wildtype (n=6)
DLX-CB1-KO (n=8)
Chance
% Correct Responses
Session
B
Errors by Trial Difficulty
% Incorrect
Difficulty
Group
WT
DLX-CB1-KO
C
P(S1) vs P(S0)
P(Sn)
Session
Group
WT
DLX-CB1-KO
P(Sn)
P(S1)
P(S0)
Chance
D
Run Time x Surface
Run time (ms)
Session
Group
WT
DLX-CB1-KO
Surface Chosen
S1
S0
E
Choice Revision x Surface
KOOKs
Session
Group
WT
DLX-CB1-KO
Surface Chosen
S1
S0
F
Choice Revision x Final Outcome
KOOKs
Session
Group
WT
DLX-CB1-KO
Final Outcome
OK
KO

Supplementary figure A. 5 - CB1 deletion from GABAergic neurons impairs EdRR performance, increases initial R1 bias, but does not impair its extinction.

All error bands represent 95% confidence intervals. (**A**) % Correct EdM responses. Dlx-CB_1-KO performed consistently more poorly than their wildtype littermates. Their mean population performance never went beyond chance level performance. (**B**) % EdM errors as a function of trial complexity. Wildtype animals displayed characteristically higher performance on level 0 trials, but Dlx-CB_1-KO performed comparably poorly on all levels, including level 0. (**C**) Probability of surface choice. Dlx-CB_1-KO displayed consistently higher P(S1) compared to wildtype littermates, but this value did decrease at a similar rate in both groups as a function of repeated EdRR training, indicating that Dlx-CB_1-KO mice were capable of revising R1 despite not being able to perform EdM. (**D**) Median run times by group and by surface. Dlx-CB_1-KO were also more sensitive to surface in terms of displaying even higher run time values on S0 compared to wildtype animals. (**E**) Mean total choice revision by group and by surface. Both groups displayed comparable phenotypes for choice revision; more in earlier EdRR sessions and favoring S1 final decisions. (**F**) Mean total choice revision by group and by outcome. Similarly, looking at choice revision as a function of final choice outcome (correct or incorrect EdM response), no significant differences were observed.


A
Classical EdM - Errors x Trial Complexity
ITI <= 10000
% Incorrect
Difficulty
Group
WT
DLX-CB1-KO
B
Classical EdM - Errors x Trial Complexity
ITI = 20000
% Incorrect
Difficulty
Group
WT
DLX-CB1-KO
C
Decision Latency x Difficulty x Outcome
Decision Latency (ms)
Difficulty
Group
WT
DLX-CB1-KO
Outcome
Correct
Incorrect
D
Run Time x Difficulty x Outcome
Run Time (ms)
Difficulty
Group
WT
DLX-CB1-KO
Outcome
Correct
Incorrect
E
Choice Revision x Diff. x Outcome
KOOKs
Difficulty
Group
WT
DLX-CB1-KO
Outcome
Correct
Incorrect

Supplementary figure A. 6 - CB1 deletion from GABAergic neurons severely impairs EdM performance.

Prior to analyzing their response to the EdRR protocol, we first characterized how Dlx-CB_1-KO mice would behave in the classical EdM task. We had predicted they may be slightly impaired, but in fact these preliminary studies revealed a more extreme phenotype than we had expected, including increased levels of cognitive behaviors in deliberative decision-making parameters. All error bands represent 95% confidence intervals. (**A**) % Incorrect responses according to difficulty. Wildtype animals (red, n=14) displayed characteristic EdM performance as a function of trial complexity, with highest scores on level 0 trials. In contrast, Dlx-CB_1-KO (green, n=15) attained only slightly, albeit significantly higher performance on level 0 and level 1 trials only. Performance on all other complexity levels did not go beyond chance. (**B**) However, when intertrial interval was increased to 20s, this improved Dlx-CB_1-KO performance on level 0 trials, reflecting an improvement putatively due to increased "passive" forgetting of interfering cognitive content. (**C**) Median decision latencies by difficulty and by outcome. Dlx-CB_1-KO displayed robustly higher decision latencies across all trial complexity levels independently of outcome and otherwise following the same phenotype seen in wildtype animals; higher decision latencies on easier trials, decreasing steadily as a function of difficulty. (**D**) Median run times by difficulty and by outcome. Wildtype animals here displayed the same characteristic run time phenotype seen in D1-CB1-KO and wildtypes; higher run time on incorrect outcome trials, higher on easier trials and decreasing as a function of difficulty. In contrast, Dlx-CB_1-KO displayed significantly higher run times with no dependence on trial complexity or outcome. (**E**) Mean total choice revision by difficulty and by outcome. Again, wildtype animals displayed the same characteristic phenotype as other groups; more choice revision on easier trials, significantly favoring correct outcomes. Dlx-CB_1-KO displayed significantly more choice revision, independently of trial complexity and outcome. These three deliberative phenotypes in Dlx-CB_1-KO indicate a surplus of active cognitive content which led us to theorize a deficit of adaptive forgetting in this mouse line.

comparison (one-way ANOVA, $F(4, 1030) = 2.2$, $p = 0.07$). In short, performances, as predicted, were better on level 0 trials, but this difference was smaller than we would have predicted, demonstrating that Dlx-CB_1-KO are in fact more impaired in the EdM task than animals with lesions to the CA1. Finally, when we increased the inter-trial interval to 20000ms, performance in Dlx-CB_1-KO mice on complexity level 0 trials only improved significantly compared to performance on the same complexity trials at ITIs of 10000ms and below (supplementary figure A.6a-b, t-test with Welch correction between complexity level 0 performance at ITI = 20000 and at ITI <= 10000, $t(22) = 3$, $p = 0.005$), constituting further evidence that the deficit in this mouse line was indeed at the level of active inhibition/adaptive forgetting, the cognitive need for which decreases as a function of increasing ITI (Al Abed et al. 2016).

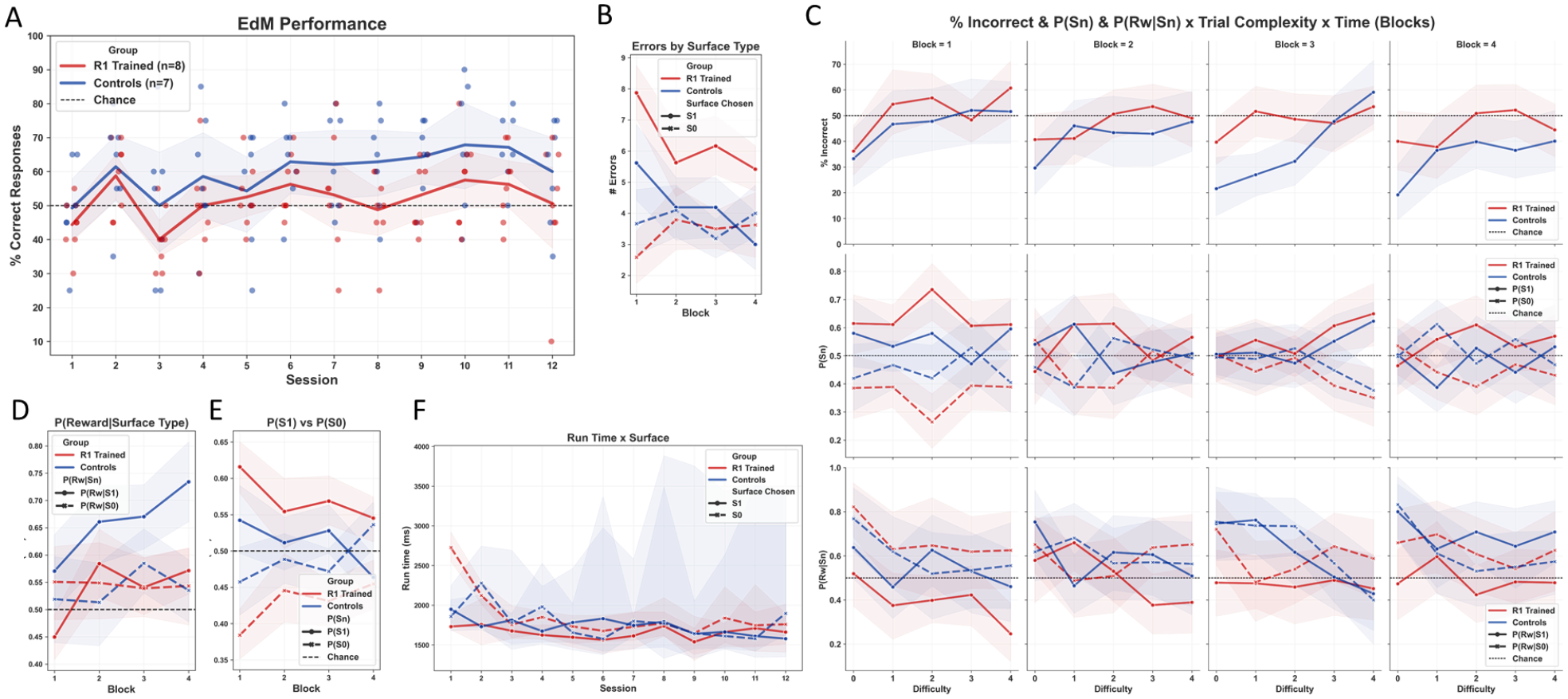
Interestingly, in these pilot experiments in the classical EdM task, we also observed that, compared to their wildtype littermates, Dlx-CB_1-KO animals displayed amplified overall cognitive activity in decision latency, run time, and choice revision behavior (one-way ANOVAs with pairwise Tukey HSD post-hoc; decision latency, $F(1, 433) = 15.3$, $p = 0.001$; run time, $F(1, 867) = 45.4$, $p = 0.001$; choice revision, $F(1, 868) = 39.1$, $p = 0.001$). However, in contrast to wildtype animals, run time was not dependent on whether or not the arm chosen was the correct one (as seen above with D_1-CB_1-KO and their wildtype littermates, run time was significantly higher, in a trial complexity-dependent manner, in wildtype animals on trials where the incorrect arm was chosen; supplementary figure A.6d; one-way ANOVA with pairwise Tukey HSD post-hoc on overall run times between 'Outcome', correct or incorrect; wildtype group, $F(1, 417) = 7.8$, $p = 0.005$; KO group, $F(1, 448) = 0.17$, $p = 0.68$; trial complexity-dependent element visible in figure). Nor did the amount of choice revision depend on trial complexity in KO animals (one-way ANOVA with pairwise Tukey HSD post-hoc revealed a significant effect of 'Difficulty' in choice revision in wildtype animals, with specifically significant effects in level 0 vs 2, 0 vs 3, and 0 vs 4, $F(4, 1045) = 8.7$, $p = 0.001$; in Dlx-CB_1-KO animals, no significant effect of 'Difficulty' was observed, $F(4, 1120) = 0.6$, $p = 0.63$). Finally, choice revision was also not more likely to be rectifying than error-inducing in Dlx-CB_1-KO mice, unlike in wildtype animals (one-way ANOVAs with pairwise Tukey HSD post-hoc on choice revision between 'Outcome' correct or incorrect; wildtype, $F(1, 418) = 20.1$, $p = 0.001$; Dlx-CB_1-KO, $F(1, 448) = 0.1$, $p = 0.72$). All of these results point towards an over-active and thereby imprecise treatment of

online cognitive content in Dlx-CB_1-KO mice, spilling over to severely affect even the spontaneous alternation equivalent level 0 complexity trials, indicating that a characteristic of their endophenotype is a dysfunction of targeted, top-down active inhibition, in turn responsible for the organizational adaptive forgetting necessary for successful EdM performance.

Deletion of CB1 receptors from GABAergic neurons of the forebrain does not impact expression of an exploration-antagonistic S-R rule: The phenotypes seen in Dlx-CB_1-KO mice above, in the hippocampus-dependent EdM task, made it all the more striking when this same mouse line displayed no phenotype in the preliminary R1 training phase of our rule revision protocol (supplementary figure A.1c). Indeed, this reinforces the idea we put forward in Stevens et al., 2022a, that the R1 tactile discrimination task is dependent upon the striatum and not the hippocampus. These R1 results also fit with recent work, mentioned above, showing that Dlx-CB_1-KO mice have no phenotypical impairment in forming direct associations (Busquets-Garcia et al. 2018). The only factor in which we observed a difference during R1 training was run time, with Dlx-CB_1-KO mice being slightly more sensitive, albeit erratically so, to the S0 surface than their wildtype littermates, giving rise to a significantly higher S0 run time when averaged across the final, fully pseudo-random sessions of R1 (supplementary figure A.1e; one-way ANOVA with pairwise Tukey HSD post-hoc within S0 run times between 'Group', $F(1, 136) = 4.3$, $p = 0.0078$).

Deletion of CB1 receptors from GABAergic neurons of the forebrain increases R1 interference during EdRR without impeding its inhibition: When both groups were moved to the EdRR phase (all subjects from both groups reached criterion in R1), we observed a significant effect of repeated training on performance in both groups. However, only wildtype animals rose above chance level EdM performance; Dlx-CB_1-KO mice, as predicted, did not (supplementary figure A.5a). Indeed, Dlx-CB_1-KO mice, but not wildtype, performed significantly below chance level during the first three sessions (t-tests between block 1 performance and chance level of 50; Dlx-CB_1-KO, $t(23) = 6.6$, $p < 0.0001$; wildtype, $t(17) = 0.8$, $p = 0.45$). In the second block, Dlx-CB_1-KO were closer to chance level but wildtype animals performed significantly better than chance level (t-tests between block 2 performance and chance level of 50; Dlx-CB_1-KO, $t(23) = 1.5$, $p = 0.14$; wildtype, $t(17) = 3.5$, $p = 0.003$). Averaging across all EdRR

sessions to look at EdRR errors by trial complexity level, we observed that R1 trained wildtype animals performed well primarily on level 0 trials, whereas R1 trained Dlx-CB_1-KO mice were at below chance performance at all trial complexity levels We also observed a trend for Dlx-CB_1-KO to have higher P(S1) and lower P(S0) values compared to wildtype animals in each session. This difference was statistically significant in the first block but not in the second (supplementary figure A.5c; one-way ANOVAs with pairwise Tukey HSD post-hoc, block 1, $F(1, 40) = 8.7$, $p = 0.005$; block 2, $F(1, 40) = 1.5$, $p = 0.23$), reflecting what is visible in supplementary figure A.5c; that, despite failure to perform beyond chance level in the EdRR task, Dlx-CB_1-KO mice did in fact inhibit R1 interference at a similar rate to their WT littermates, albeit beginning from a higher starting point. In terms of run time, we observed a similar and significant trend as in the R1 phase for the Dlx-CB_1-KO mice to be more sensitive to S0, with even larger run times on this surface compared to WT mice (two-way ANOVA, 'Group' and 'Surface', $F(1, 164) = 4.8$, $p = 0.03$, with Tukey HSD post-hoc on 'Group*Surface' interaction revealing a significant KO S0 vs WT S0 effect, $p = 0.001$). When we analyzed choice revision (supplementary figure A.5e-f), we observed that, in contrast to the classical EdM pilot we had conducted (supplementary figure A.6e), Dlx-CB_1-KO animals did not display this behavior more than wildtype animals in the EdRR task (one-way ANOVA, $F(1, 166) = 1.29$, $p = 0.4$). This may be explained by the fact, observed in our characterization of the EdRR paradigm (figure 7e-f), that prior R1 learning in itself increases choice revision behavior, which may occlude a prior learning-independent trend towards increased choice revision specifically in the Dlx-CB_1-KO. We did, however, observe that in the second block (sessions 4-6) Dlx-CB_1-KO but not wildtype mice were still revising their choice more towards S1 than towards S0 (block 2, one-way ANOVAs within 'Group' between 'Surface' with pairwise Tukey HSD post-hoc; WT, $F(1, 34) = 1.4$, $p = 0.71$; Dlx-CB_1-KO, $F(1, 82) = 9.6$, $p = 0.003$). This result coincides with the fact that in block 2, wildtype animals began to revise their choice more often in a rectifying than in an error-inducing manner, to a greater extent than Dlx-CB_1-KO mice, though this did not reach statistical significance. On this last point, we again draw attention to the preliminary classical EdM experiment we conducted with the Dlx-CB_1-KO line, during which we observed that KO animals, despite performing significantly more choice revision, did so in a way that was no more likely to give rise to a correct than an incorrect EdM response (supplementary figure A.6e).


A
EdM Performance
Group
R1 Trained (n=8)
Controls (n=7)
Chance
% Correct Responses
Session
B
Errors by Surface Type
Group
R1 Trained
Controls
Surface Chosen
S1
S0
Errors
Block
C
% Incorrect & P(Sn) & P(Rw|Sn) x Trial Complexity x Time (Blocks)
Block = 1
Block = 2
Block = 3
Block = 4
% Incorrect
P(Sn)
P(Rw|Sn)
R1 Trained
Controls
P(S1)
P(S0)
P(Rw|S1)
P(Rw|S0)
Chance
Difficulty
D
P(Reward|Surface Type)
Group
R1 Trained
Controls
P(Rw|Sn)
P(Rw|S1)
P(Rw|S0)
Block
E
P(S1) vs P(S0)
Group
R1 Trained
Controls
P(Sn)
P(S1)
P(S0)
Chance
Block
F
Run Time x Surface
Group
R1 Trained
Controls
Surface Chosen
S1
S0
Run time (ms)
Session

Supplementary figure A. 7 - Aged mice displayed robust EdM deficit in conditions of R1 rule revision compared to controls.

Aged R1 trained animals (red, n=8) displayed robust R1 bias but, unlike young adult R1 trained mice, also performed more poorly than controls (blue, n=7) in EdM across all sessions of EdRR. Error bands represent 95% confidence intervals. Vertical spaces between error bands indicate statistical significance (for detailed statistical analyses, see main text). (**A**) % Correct EdM responses. As with young adult mice, both R1 trained and control animals performed poorly in initial sessions. With repeated training, controls, especially in blocks 3 + 4, improved their performance slightly, whereas the R1 trained population did not. Averaged across all sessions, controls thus performed significantly better. (**B**) Errors by surface type by block. R1 trained animals were highly biased in their errors, especially in block 1. This bias decreased rapidly but remained significant in later blocks also. (**C**) top row, % Incorrect EdM responses by difficulty and by block. Control animals performed significantly better on level 0 trials in all blocks, with their performances on all levels, but especially 0, globally improving over repeated EdRR training. R1 trained animals performed slightly better on level 0 trials, significantly so in block 1 and averaged across all blocks, but did not perform at higher than chance level on any other trial complexity level. (**C**) middle row + (**E**) P(Sn) by difficult and by block. Control animals displayed highly erratic behavior in terms of surface choice probability, nevertheless with a strong tendency, significant when averaged across all blocks to have higher P(S1) values on level 4 trials, similar to what had been observed in young adult controls. R1 trained animals displayed globally higher P(S1) values, especially in block 1, but this too was more erratic than what had been observed in young adult mice. (**C**) bottom row + (**D**) P(Rw|Sn) by difficulty and by block. In aged controls, as with young adults, P(Rw|S1) was globally higher than P(Rw|S0), which could, once again, explain their higher P(S1) values on level 4 trials. R1 trained animals had higher P(Rw|S0) values than P(Rw|S1) values but in block 1 only, after which P(Rw|Sn) values fluctuated somewhat erratically. (**F**) Median run times by surface by session. Run times in aged control animals displayed very high variability. R1 trained animals displayed characteristic higher S0 run times in initial sessions as well as slightly but reliably lower S1 run times across all sessions. Forthcoming work from our lab shows that aged mice have a deficit in adaptive forgetting. This phenotype combined with the persistent resistance they demonstrated to sustained R1 expression during R1 training may account for the combination seen here, i.e. robust extinction of R1 bias coupled, counter-intuitively, with a global deficit in EdM performance. If aged mice are impaired in active inhibition of interfering cognitive content in classical EdM conditions, then the addition of the extra R1 cognitive content may serve simply to exacerbate this, albeit indirectly. Further investigation is required into these open questions.

A.2 Ageing approach

The physiological ageing process is known to have important negative impacts on inhibitory cognitive flexibility (Coxon et al. 2012). Our own lab has published important work in the area of age-related decline of declarative memory capacity using both the EdM and other radial maze-based tasks (Marighetto et al. 1999; Marighetto, Brayda-Bruno, and Etchamendy 2011). More specifically, with respect to the CB_1-based genetic approaches we employed above, it has been demonstrated that levels of CB_1 decrease with age, thereby weakening certain functions of the endocannabinoid system (Bilkei-Gorzo 2012; Bilkei-Gorzo et al. 2017; Albayram et al. 2011). Taking all this together, we had predicted that, similar to D_1-CB_1-KO mice, aged mice would be impaired in expression, but not acquisition, of the R1 tactile discrimination rule. Such an impairment is precisely what we observed (supplementary figure A.1d; see also Stevens et al., 2022a). Aged mice required longer R1 training than young mice to reach criterion R1 performance, and some animals, even with additional training, still did not reach this level supplementary figure A.1d zoombox). Knowing from previous work from our lab that aged mice are impaired in the classical EdM task (putatively due to a decline in the active inhibition/adaptive forgetting necessary for organizing cognitive content), and now knowing they were also impaired in expression of R1 (putatively due to a decline in the active inhibitory cognitive flexibility necessary to overcome spontaneous exploratory behavior), no self-evident hypothesis was forth-coming regarding how these phenotypes would combine when brought together in the context of the EdRR environment. Therefore, with respect to the exact nature of R1 interference in the EdRR phase, we proceeded without a clear hypothesis of how performance in the EdRR task may differ in aged compared to young mice. In these experiments, we again used a control group of aged littermates who were, as with the original experiment in young C57Bl6/J mice, rewarded on every trial during R1 training, regardless of whether they chose S1 or S0.

A.2.1 EdM performance negatively impacted by R1 interference in aged mice.

Having eliminated the three below criterion aged mice, 8 R1 trained aged animals and all 7 aged controls were moved to the EdRR environment. Here, we trained both groups,

as per the original experiment (figure 2a), for 12 sessions in order to track the evolution of the impact of prior R1 learning (supplementary figure A.7a). Most interestingly, compared to our observations in young mice, aged R1 trained mice displayed a robust and persistent impairment in the EdM performance itself compared to controls (supplementary figure A.7a, one-way ANOVA averaged across all sessions with pairwise Tukey HSD post-hoc and unbiased Cohen effect size, $F(1, 178) = 15.6$, $p = 0.001$, $d = 0.59$). Additionally, while EdM performance seemed to improve slightly as a function of repeated training in controls, this was not the case in the R1 trained group, though this evolution achieved significance in neither group when corrected for lack of sphericity (repeated measures ANOVA within 'Session' with Greenhouse-Geisser correction for $\varepsilon < 1$; R1 trained group, $F(11, 77) = 1.6$, uncorrected $p = 0.1$, GG corrected $p = 0.2$; control group, $F(11, 66) = 2.2$, uncorrected $p = 0.026$, GG corrected $p = 0.1$). With respect to the nature of the errors made, as was the case in young mice (figure 2b), aged R1 trained mice made significantly more errors on S1 than S0 arms (supplementary figure A.7b, sessions averaged across blocks of 3; mixed ANOVA significant interaction between 'Group' within 'Surface', $F(1, 13) = 4.7$, $p = 0.049$; Tukey HSD post-hoc on interaction revealed significant differences between R1 trained S1 and S0 errors, $p = 0.001$; between R1 trained S1 and control S1 errors, $p = 0.001$; but no significant difference between control S1 and S0 errors, $p = 0.5$). When we looked at the number of errors committed as a function of repeated training and trial complexity (supplementary figure A.7c, top row), we observed that performance in control animals, but not in R1 trained animals, tended to improve with repeated training (each column, left to right, represents a consecutive block; one-way ANOVA with pairwise Tukey HSD post-hoc within each group revealed a slight but significant effect of 'Block' on performance only in control animals, $F(3, 416) = 3.4$, $p = 0.018$; R1 trained, $F(3, 476) = 1.2$, $p = 0.26$). There was also a significant effect of trial difficulty in both groups, the size of which was nevertheless greater in controls (one-way ANOVA with pairwise Tukey HSD post-hoc, R1 trained, $F(4, 475) = 4.1$, $p = 0.003$; controls, $F(4, 415) = 8.4$, $p < 0.0001$).

In terms of surface choice probability, 'P(Sn)', we did not observe the same direct linear effect of trial complexity as we had in young animals (see figure 4a, middle row), but the general trend for P(S1) to be lower in R1 trained animals at level 0 still emerged clearly over time (supplementary figure A.7c, middle row). Averaged across all blocks,

there was a significant effect of trial difficulty on P(Sn) values in R1 trained animals but not in controls (one-way ANOVA with pairwise Tukey HSD post-hoc, R1 trained, $F(4, 475) = 3$, $p = 0.019$; controls, $F(4, 415) = 1.3$, $p = 0.28$). Again averaged across all blocks, there was a trend, similar to what we had observed in young control mice, for P(S1) to be greater than P(S0) on complexity level 4 trials.

P(Rw|Sn) reflected similar trends to those seen in young mice (see figure 4a, bottom row), with aged R1 trained animals significantly more likely to obtain a reward when they chose S0 as opposed to S1 (two-way ANOVA on interaction 'Group*P(Rw|Sn)', averaged across all blocks, $F(1, 1668) = 16.2$, $p < 0.0001$; pairwise Tukey HSD post-hocs, R1 P(Rw|S0) vs R1 P(Rw|S1), $p = 0.001$; Ctrl P(Rw|S0) vs Ctrl P(Rw|S1), $p = 0.9$). As with the results seen in young mice, the relation of P(Rw|Sn) results to rule revision come into strongest focus when viewed side by side with the P(Sn) results (supplementary figure A.7d-e). Here we again see that P(Sn) in R1 trained mice does not evolve to match their P(Rw|Sn) values. As we saw above, young control animals tended to develop a preference for P(S1), which was consistent with the fact that their P(Rw|S1) values were reliably higher than their P(Rw|S0) values (see figure 4a-b). In aged control mice also, P(Rw|S1) was reliably higher than P(Rw|S0) (supplementary figure A.7d), yet their P(S1) was erratic, notably going below P(S0) during the final block (supplementary figure A.7e). In aged R1 trained mice, mean block-by-block P(Rw|S1) and P(Rw|S0) values reached similar values by the second block, much earlier than with young R1 trained mice (see figure 5a). One possibility, inspired in part by the results seen when we placed below R1 criterion D_1-CB_1-KO mice in the EdRR task, is that the high level of resistance to R1 expression manifest during R1 training led them, in the EdRR phase, to return sooner and more strongly to exploration based behavior than young R1 trained animals had. If this were the case, then the age-related decrease in cognitive flexibility underpinning a "rigid" drive to explore, coupled paradoxically with their age-related inherent limitation with respect to the cognitive requirements of the EdM task, would translate into the principal net behavioral result we observed in the aged R1 trained population; a compounding and exacerbation of the usual age-related classical EdM impairment.

Finally, we also looked at run time behavior in aged mice during the EdRR phase. As one might expect, run times were globally higher in aged compared to young mice

(overall median run time during EdRR, independent of group, trial difficulty, and surface; aged mice, 1782.5ms; young mice, 1291ms). This inherent reduction in locomotor speed with age may therefore have served to occlude some of the differential cognitive effects we would expect to see following R1 training. Nevertheless, run times were still significantly higher on S0 compared to S1 in aged R1 trained animals during the first block of EdRR training, but not in controls (supplementary figure A.7f, one-way ANOVA with pairwise Tukey HSD post-hoc, $F(1, 46) = 24$, $p = 0.001$; controls $F(1, 40) = 0.3$, $p = 0.58$). Furthermore, similar to what we had observed in younger mice, even beyond this significant first session difference, median S1 run times remained robustly lower than median S0 run times in aged R1 trained animals, giving rise to a significant effect averaged across all 4 blocks of repeated EdRR training (one-way ANOVA with pairwise Tukey HSD post-hoc, $F(1, 190) = 11.9$, $p = 0.001$; controls, $F(1, 166) = 0.9$, $p = 0.34$). (We note that, in terms of run time, control animals displayed highly erratic behavior, visible in the large error band variance in supplementary figure A.7f. For this reason, we restricted our run time analyses to between surface comparisons within the R1 trained group.)

Appendix Bibliography

Al Abed, Alice Shaam, Azza Sellami, Laurent Brayda-Bruno, Valérie Lamothe, Xavier Noguès, Mylène Potier, Catherine Bennetau-Pelissero, and Aline Marighetto. 2016. "Estradiol Enhances Retention but Not Organization of Hippocampus-Dependent Memory in Intact Male Mice." *Psychoneuroendocrinology* 69 (July): 77–89. https://doi.org/10.1016/j.psyneuen.2016.03.014.

Albayram, O., J. Alferink, J. Pitsch, A. Piyanova, K. Neitzert, K. Poppensieker, D. Mauer, et al. 2011. "Role of CB1 Cannabinoid Receptors on GABAergic Neurons in Brain Aging." *Proceedings of the National Academy of Sciences* 108 (27): 11256–61. https://doi.org/10.1073/pnas.1016442108.

Albayram, O., Stefan Passlick, Andras Bilkei-Gorzo, Andreas Zimmer, and Christian Steinhäuser. 2016. "Physiological Impact of CB1 Receptor Expression by Hippocampal GABAergic Interneurons." *Pflügers Archiv - European Journal of Physiology* 468 (4): 727–37. https://doi.org/10.1007/s00424-015-1782-5.

Bilkei-Gorzo, Andras. 2012. "The Endocannabinoid System in Normal and Pathological Brain Ageing." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1607): 3326–41. https://doi.org/10.1098/rstb.2011.0388.

Bilkei-Gorzo, Andras, Onder Albayram, Astrid Draffehn, Kerstin Michel, Anastasia Piyanova, Hannah Oppenheimer, Mona Dvir-Ginzberg, et al. 2017. "A Chronic Low Dose of Δ9-Tetrahydrocannabinol (THC) Restores Cognitive Function in Old Mice." *Nature Medicine* 23 (6): 782–87. https://doi.org/10.1038/nm.4311.

Busquets-Garcia, Arnau, José F. Oliveira da Cruz, Geoffrey Terral, Antonio C. Pagano Zottola, Edgar Soria-Gómez, Andrea Contini, Hugo Martin, et al. 2018. "Hippocampal CB1 Receptors Control Incidental Associations." *Neuron* 99 (6): 1247-1259.e7. https://doi.org/10.1016/j.neuron.2018.08.014.

Coxon, James P., Annouchka Van Impe, Nicole Wenderoth, and Stephan P. Swinnen. 2012. "Aging and Inhibitory Control of Action: Cortico-Subthalamic Connection Strength Predicts Stopping Performance." *Journal of Neuroscience* 32 (24): 8401–12. https://doi.org/10.1523/JNEUROSCI.6360-11.2012.

Graybiel, Ann M. 1998. "The Basal Ganglia and Chunking of Action Repertoires." *Neurobiology of Learning and Memory* 70 (1–2): 119–36. https://doi.org/10.1006/nlme.1998.3843.

Hitoshi, Niwa, Yamamura Ken-ichi, and Miyazaki Jun-ichi. 1991. "Efficient Selection for High-Expression Transfectants with a Novel Eukaryotic Vector." *Gene* 108 (2): 193–99. https://doi.org/10.1016/0378-1119(91)90434-D.

Jin, Xin, Fatuel Tecuapetla, and Rui M. Costa. 2014. "Basal Ganglia Subcircuits Distinctively Encode the Parsing and Concatenation of Action Sequences." *Nature Neuroscience* 17 (3): 423–30. https://doi.org/10.1038/nn.3632.

Marighetto, Aline, Laurent Brayda-Bruno, and Nicole Etchamendy. 2011. "Studying the Impact of Aging on Memory Systems: Contribution of Two Behavioral Models in the Mouse." In *Behavioral Neurobiology of Aging*, edited by Marie-Christine Pardon and Mark W. Bondi, 10:67–89. Current Topics in Behavioral Neurosciences. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/7854_2011_151.

Marighetto, Aline, Nicole Etchamendy, Khalid Touzani, Cedric Cortes Torrea, Benjamin K. Yee, John Nicholas P. Rawlins, Robert Jaffard, and Marighetto, Aline. 1999. "Knowing Which and Knowing What: A Potential Mouse Model for Age-Related Human Declarative Memory Decline: Mouse Model for Memory Decline." *European Journal of Neuroscience* 11 (9): 3312–22. https://doi.org/10.1046/j.1460-9568.1999.00741.x.

Soria-Gomez, Edgar, Antonio C. Pagano Zottola, Yamuna Mariani, Tifany Desprez, Massimo Barresi, Itziar Bonilla-del Río, Carolina Muguruza, et al. 2021. "Subcellular Specificity of Cannabinoid Effects in Striatonigral Circuits." *Neuron* 109 (9): 1513-1526.e11. https://doi.org/10.1016/j.neuron.2021.03.007.

General Conclusion & Perspectives

The novel two-fold model of indoctrination-like learning and everyday-like rule revision, which constitutes the principal component of the thesis project presented here, raises more questions than it answers. Its strength, however, is in providing the means for investigating those questions. Here, we have shown what has been discovered in pursuing just two such avenues of research: using conditional deletion and viral re-expression of specific populations of CB_1 receptors as a technique for isolating and modulating the contributions to behavior of certain memory systems but not others, and; using aged mice to study how cognitive decline impacts the availability of cognitive resources necessary for organizational memory, especially in environments that demand revision of a prior rule. As an extension to the research already conducted here, immediate future work will continue to characterize the specific role of the endocannabinoid system in both everyday-like memory and everyday-like rule revision, especially since this system has already demonstrated its promising potential as a therapeutic target for age-related decline in declarative memory performance (Bilkei-Gorzo et al., 2017). More precisely, we wish to conduct viral CB_1 re-expression studies in the Dlx-CB_1-KO mouse line in order to answer the question of whether their apparent lack of inhibitory control, visible in the combination of increased deliberative behaviors yet inability to successfully perform EdM or EdRR, is due to CB_1 loss specifically in the prefrontal cortex or to CB_1 loss locally in the hippocampus. Such viral re-expression experiments had in fact been begun at precisely the moment the first lockdown of the COVID health crisis was announced, incurring the regrettable but necessary loss of all our operated experimental animals. Along similar lines, we also aim to conduct viral CB_1 over-expression experiments in aged mice, separately in the dorsal striatum, relative to their poor performance in inhibiting the exploratory drive during R1, and in the prefrontal cortex or hippocampus, relative to their poor performance in EdM and EdRR. These experiments represent pre-clinical studies which have the potential to rapidly contribute to human clinical research into age-related memory decline.

Another future experimental avenue will make use of the optogenetic set up recently developed by our team for targeted control of discrete neuronal populations *in vivo* in freely behaving animals in the radial maze. The modular scope of the radial maze will allow us, for example, to monitor the effects of inhibition or activation of specific EdM

or EdRR task relevant circuits in a pair specific manner, thus creating a situation of within-subject control. To take just one example as illustration, the medial prefrontal cortex could be inhibited on one out of three pairs during EdRR, in which case we would predict that performance and R1 bias would be increased on this pair, even on level 0 trials, compared to the other two pairs (hypothesis based on the importance of top-down mPFC mediated inhibitory control for successful performance of the EdM and EdRR tasks). Similarly, through targeted activation or inhibition of amygdalar or dorso-striatal activity, we will be able to directly test hypotheses relating to the putative contributions of these regions to, for example, run time and choice revision behaviors.

Finally, with respect to experiments which we already have the technical capacity to undertake within the team, we are currently in the process of setting up a fiberphotometry system for use *in vivo* in freely behaving mice in the radial maze. This technique comes with vast potential in terms of monitoring brain activity during specific discrete behavioral episodes, such as slower run times on incorrect choices during EdM and on S0 arms during R1 and EdRR, or in the instants leading up to, during, and immediately following physical choice revisions in both tasks. The development of tools of analysis during the present thesis project, enabling the identification and temporal isolation of such discrete episodes, provides the possibility of relating monitored brain activity to such events in a classical and powerful peri-event manner. This combination of discrete behavior identification and *in vivo* neural activity monitoring is an extremely exciting prospect for future research in our team.

We have already begun work on expanding the scope of our animal model of EdRR by developing a virtual version for human subjects. Such a virtual model has already been published for the classical EdM task, and we believe an EdRR version would be of particular interest in today's world, given the ever more evident fact that confirmation bias is a phenomenon we cannot afford to have only partial understanding of. Already, the simple fact of having succeeded in producing comparable and robust effects in mice, by itself substantially alters our previous best understanding of myside confirmation bias. Furthermore, specific phenotypes for confirmation bias have already been related, as mentioned above in the introduction to "Investigating hallmarks of 'myside' confirmation bias in a novel mouse model of everyday-like rule revision," to several physiological and pathological psychological conditions. Consequently, having a virtual

model capable of eliciting myside confirmation bias-like behavior could become a tool not only for studying the phenomenon in clinical patients but also for discovering how best to cognitively guide them in overcoming it.

Finally, the power of computational modeling for gaining deeper insights into behavior and for providing the means to develop and test functional hypotheses cannot be overstated. For this reason, we have already established collaborative links with members of the Mnemosyne Team at the Institut des Maladies Neurodégénératives (IMN) here in Bordeaux with a view to projects that will exploit the power of computational models towards deepening our understanding of the various behaviors elicited by both the EdM and EdRR tasks. Such complex tasks, which mobilize a multiplicity of interacting memory systems, constitute exciting challenges for the computational community and, in turn, computational models offer powerful means for generating predictions (i.e. over-activation of circuit X within a model will lead to phenotype Y) which can then be tested empirically using techniques such as those mentioned above; modulatory transgenic manipulations, *in vivo* optogenetics, and *in vivo* monitoring of neural activity.

The important point we wish to retain and communicate in final conclusion to this doctoral research is that all of the above possibilities will have been enabled by initial, attentive, detailed, and meticulous observation and description of animal responses to behavioral tasks developed as much as possible to reflect real world situations. Just how essential this research dimension is to our capacity, as neuroscientists, to say something meaningful about brain function is a message which has still not fully spread throughout the community, but which we are determined to add our voice to (Genzel, 2021; Krakauer et al., 2017; Niv, 2021; Staddon & Simmelhag, 1971).

General conclusion references:

Bilkei-Gorzo, A., Albayram, O., Draffehn, A., Michel, K., Piyanova, A., Oppenheimer, H., Dvir-Ginzberg, M., Rácz, I., Ulas, T., Imbeault, S., Bab, I., Schultze, J. L., & Zimmer, A. (2017). A chronic low dose of Δ9-tetrahydrocannabinol (THC) restores cognitive function in old mice. *Nature Medicine*, *23*(6), 782–787. https://doi.org/10.1038/nm.4311

Genzel, L. (2021). How to Control Behavioral Studies for Rodents—Don't Project Human Thoughts onto Them. *ENeuro*, *8*(1). https://doi.org/10.1523/ENEURO.0456-20.2021

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, *135*(5), 601. https://doi.org/10.1037/bne0000471

Staddon, J. E., & Simmelhag, V. L. (1971). The "supersitition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psychological Review*, *78*(1), 3–43. https://doi.org/10.1037/h0030305

Detailed Table of Contents

Neural and cognitive bases of confirmation bias-induced interference in declarative memory performance.

Confirmation bias is a well-described and ubiquitous cognitive behavior whereby novel information from the environment is over-valued when it confirms and under-valued when it disconfirms previously consolidated cognitive content (e.g. beliefs, learned associations, etc.). The maladaptive responses this phenomenon can give rise to are implicated in social problems such as the spread of "fake news" and vary according to both contextual complexity and the mental state of the subject.

Nevertheless, very little research has been dedicated to understanding the neural mechanisms or evolution underpinning this spontaneous human cognitive response to novel information. Thus, we designed a mouse model for confirmation bias-like behavior, enabling exploration of its cognitive and neurobiological underpinnings and their evolution. Our model is based on a cognitive level definition of the phenomenon; over-valuation of novel environmental elements which confirm and under-valuation of novel environmental elements which disconfirm a previously consolidated cognitive content. Our results to this point (using a two-task, two-context radial maze protocol) show a strong bias effect which is observable as a deviation in the performance of a classical declarative memory task, the persistence of which is trial-complexity dependent.

Detailed behavioral analysis has enabled us to identify several more basic cognitive components impacting the bias effect, such as adaptive forgetting and the exploration/exploitation balance. These cognitive components have been identified with specific neural circuits whose activity is susceptible to intervention and/or monitoring in freely moving task-performing animals. They are also implicated in many psychiatric conditions (depression, schizophrenia, etc.) making of this model a novel tool for pre-clinical research of which we are developing a human version for clinical research and a computational version for formulating and testing predictions.